THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Estimation and Application of Bayesian Hawkes Process Models

THE UNIVERSITY
of EDINBURGH

Isabella Deutsch

Doctor of Philosophy
University of Edinburgh
2023

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

(*Isabella Deutsch*)

To Gerti and Walter

Abstract

In this thesis, we examine various facets of Bayesian approaches to Hawkes Processes. Hawkes Processes are a flexible class of point processes that are used to model events that occur in clusters or bursts, as classic Hawkes processes capture the self-exciting behaviour where one event makes future events more likely. While they are popular in the earthquake literature, they are also successfully used in other applications, such as crime, email or Twitter messaging patterns, or tradings on the stock market.

A variety of estimation procedures, both in the frequentist and Bayesian domains, exist to estimate the parameters of the Hawkes process. The goal of this thesis is to enable and improve parameter estimation for different scenarios, such as missing data and inhibition. We use these findings to apply Hawkes processes to product sales analysis, specifically to identify product cannibalisation, and to model data from a group chat setting.

We address issues in parameter estimation in the excitation-only case when data from a Hawkes process is missing. This can severely bias the learning of the Hawkes process parameters. As such, we develop a novel estimation approach based on Approximate Bayesian Computation.

We then consider an extension of the Hawkes process which incorporates inhibition, where events can decrease the intensity function. This leads to additional complexities in the estimation procedure. We resolve challenges regarding the integration of the intensity function and introduce a new, less restrictive condition for stability as existing conditions are unnecessarily strict under inhibition.

Based on these findings, we use the multivariate Hawkes process to model product sales. In particular, we are interested in product cannibalisation, which refers to the decrease in the sales of one product due to competition from another product. We examine this phenomenon in a wholesale data set provided by an international company using a multivariate Hawkes process with inhibition. For this, we design a dimension-independent prior for inhibition based on a reparametrisation.

Finally, we propose an extension to the classic multivariate Hawkes model, which permits different influences for immigrant and triggered events subject to the latent branching structure. We showcase this extended model on data from a group chat.

Lay Summary

Many phenomena of interest can be described as events happening in time. On a big scale, we record when earthquakes happen in a particular region or when a football team wins a big tournament. However, events that we are interested in also happen much more often than big earthquakes or success in sports. We can monitor when calls arrive in a call centre or products get ordered, where each call or order constitutes one event. Event series are also found in a biological context. For example, neurologists can record when individual neurons in the brain fire.

Sometimes events tend to happen in clusters, where one event makes others more likely. This is called *excitation*. Examples of this are the aforementioned earthquakes, where large earthquakes are typically followed by many smaller aftershocks. However, the contrary is also possible, where the occurrence of one event makes it less likely for others to follow suit. This is *inhibition*. For example, we know that when a neuron has just fired it is less likely to fire again immediately. Instead, it has a cool-off period where events become rather unlikely.

When the arrival of any such events is recorded, the resulting event series can be analysed in a statistical manner. Such a statistical approach can help us understand why events tend to happen when they happen. Maybe they are governed by a seasonal trend. If we examine the event series of ice cream sales, more events will happen in the summer months. Maybe they follow the excitation or inhibition behaviour described above. Furthermore, a statistical approach can also help us predict when events might happen in the future. For a call centre, for example, it would be important to know how many calls will roughly reach them in a given hour such that they can schedule the appropriate number of staff.

One such statistical way to analyse the patterns of arriving events is called a *Hawkes process*. This tool is particularly useful when events occur in clusters or bursts, where one event makes other events more likely. A variety of event series can be examined with Hawkes processes. Their applications range from finance, crime, and emails, to tweets on Twitter/X and tradings on the stock market.

While Hawkes processes are a useful and versatile tool, they can be cumbersome to work with from a statistical point of view. For example, it can take quite long for a computer to complete all the steps needed to produce an analysis of an event series or a prediction of what might happen next. Therefore, the goal of this thesis is to make it easier to use Hawkes processes in some specific cases. In addition, this thesis showcases some new areas where Hawkes processes can be used successfully.

In Chapter 3 we look at a setting where events, that did indeed happen, are missing from our recorded data. Such missing data is plausible when working

with real-world data. For example, Tucker *et al.* (2019) examine a case where physical index cards, on which event records were kept, were lost. When data is missing, many classic Hawkes process approaches become inadequate as they assume that the data is collected in full. To circumvent this, we propose a tailor-made method that can be used in this case as it accounts for the fact that data is missing.

Chapter 4 focuses on Hawkes processes with both excitation and inhibition. While the addition of inhibition may seem like a small extension, it requires specific considerations such that it can be used effectively. For example, some practical mathematical shortcuts developed for the excitation-only Hawkes processes cannot be applied. Hence, we focus on improving some aspects of the Hawkes process with excitation and inhibition such that it becomes easier to use in any application.

Chapter 5 showcases Hawkes processes with inhibition in a new setting. This research is rooted in our collaboration with an international company. As they sell goods to individual customers as well as wholesale customers, they are interested in how products sell together. With a variety of similar products available, customers need to make purchasing decisions and prioritise buying some products over others. For example, a consumer may choose to buy one particular good instead of a different (yet similarly designed) one. Such an effect is known as *product cannibalisation*. We use Hawkes processes with inhibition to better understand product cannibalisation.

Finally, Chapter 6 explores how Hawkes processes can be used to analyse data from a group chat. For this line of work, we specifically collected a data set consisting of timestamps indicating when messages were sent in a group chat and who sent them. We then propose an extension of the classic Hawkes process to analyse this kind of data. Our approach helps uncover some patterns found in group chat message data, for example, one person is more likely to answer to some people than others.

Altogether, this thesis allows us to explore different aspects of the Hawkes process in a variety of applications. It showcases the applicability and versatility of this approach when it comes to analysing series of events.

Acknowledgements

I could not have undertaken this journey without my supervisor Gordon Ross. He has been a fantastic source of wisdom and has helped me develop professionally in so many ways. I am particularly thankful for his consistent willingness to consider and value my opinions and ideas. I also appreciated that any and all questions were welcomed in our discussions. His feedback and mentorship have greatly improved the quality of all my work, including this thesis. It's been a pleasure working together on so many interesting projects!

I am deeply grateful to our industry sponsor for making this PhD project possible. It was amazing working with all of you. In particular, Michael and Alex thank you for your guidance and for sharing your knowledge throughout this project. I'm so glad we got to meet in person eventually!

The completion of my degree would not have been possible without my examiners, Finn Lindgren and Ioanna Manolopoulou. Thank you for all the time and effort you invested, which resulted in a rewarding viva experience. I thoroughly enjoyed the stimulating, yet challenging, discussions I got to have with you about my research.

I am thankful to the University of Edinburgh for providing an environment for students to flourish. A big thanks for many opportunities to tutor, to grow, and to find and build community. It's an exceptional place in an exceptional city.

My degree was built on a solid statistical foundation thanks to the University of Oxford and the University of Vienna. Geoff Nicholls kindly supervised my Master's thesis and, with that, taught me an immense amount about research and Bayesian statistics. Andreas Baierl gave me the best introduction to programming and a general scientific approach I could have wished for when supervising my Bachelor's thesis.

This degree would not have been the same without a group of like-minded friends, The Usual Suspects: Andrew B, Augi J, Cob JRJB, Jamie B, Jon E, Josh F, Linden DH, Mary L. Andrew, our coffee walks brightened up so many of my days and our Highland adventure was truly memorable. Cob, thank you so much for all our side projects and corresponding shared Google Docs. We've really gone to great `library("lenght")`. Josh, there is no one I would have rather been a student rep with for two years than you! Linden, what our research topics lack in proximity, we make up for in similarity in music taste.

I was fortunate enough to meet so many amazing people throughout my degree. Jonna R, I always appreciate our tea times (virtual or otherwise). Thank you for your kindness and honesty. Torben S, our like-mindedness in statistics, language, and coffee, always makes me look forward to our next catch-up. Fur-

thermore, I had the pleasure of co-founding the Piscopia Initiative with Tiffany V and Mary L in 2019. Thank you for being my co-conspirators – I am so proud of what we achieved! In addition, I want to thank Déborah S and Raiha B for enlightening conversations about Hawkes processes.

My time at the University of Edinburgh was enhanced through serving as a trustee at Edinburgh University Students' Association. I have many talented and dedicated people to thank for amazing 2+ years there, such as Steve H, Sarah P, and Ellen MacR to name a few. It was an honour working with all of you!

A brief stint outside of the academic world brought me to Concept Ventures. A massive thank you to the whole team for welcoming me with open arms and never-ending enthusiasm. Oliver K, Reece C and Jeff C, thank you for taking a punt and allowing me to make the most of my three months with you.

There are a few more people who were essential in making Scotland my home away from home: Jamie McA, I am delighted to be able to count you as one of my closest friends. Your insights and honesty are most appreciated. Mike C, it was always great to have you in Edinburgh and I can't wait to visit you! Philip A, it's a pleasure spending evenings together, be it with food, wine, or both! Astrid C, Crick C, Jutta S, and Myriam B, thank you for welcoming me so warmly into your lives.

These acknowledgements would not be complete without mentioning the ones along the way and/or back home. Sophie Z thank you for many, many years of friendship, video calls, book club, and everything in between! Look how far we have come – I'll be your cheerleader wherever we go! Corina R and Andi L, Sofia C (now in Edinburgh as well!), Timi T and Xaver S, Consti L and Manu M, Sebastian R, Elvira W, Carina E, Thomas W and Patricia W – thank you all!

This degree would not have been half as much fun without my wonderful partner Conor. Your continuous support and encouragement, alongside your enthusiasm and calmness, were cornerstones of my last years. Thank you for delivering many cups of coffee and words of optimism at just the right time. I loved co-woking with you and our cat Lily, who routinely insisted that my keyboard was the perfect spot for a nap. Thank you for making sure I had a rich life outside of my degree, and for sharing it with me.

Finally, there are not enough words to thank my parents for their never-ending, unwavering support since Day 0. Whatever adventure I am embarking on, I can be certain of your love and reassurance through thick and thin. I know that I can always, always rely on you.

Ami and Api, thank you for giving me roots and wings.

Contents

Chapter 1

Introduction

In this thesis, we explore Hawkes processes (Hawkes, 1971) from a Bayesian point of view. Hawkes processes are a class of point processes used to model event data when events occur in clusters or bursts. Hawkes processes have a plethora of applications. They are successfully employed in finance (Laub *et al.*, 2015; Rambaldi *et al.*, 2017), cybersecurity (Bessy-Roland *et al.*, 2021), to model crime and terror (Mohler, 2013; Shelton *et al.*, 2018; Tucker *et al.*, 2019), or retweets on Twitter/X (Mei and Eisner, 2017; Rizoiu *et al.*, 2017). They are also popular in the earthquake literature as they excel at capturing the *self-exciting* behaviour of earthquakes (Ogata, 1988) where one, typically large, earthquake is followed by others of various sizes in the Epidemic Type Aftershock-Sequences (ETAS) model. Whenever events happen in close succession of each other, Hawkes processes may be useful. In this thesis, we explore a variety of aspects relating to Hawkes processes: missing data, inhibition, application to product cannibalisation, and a new parameter structure based on the branching structure showcased on data from a group chat.

First, we consider the estimation of the Hawkes process when events are missing. When working with data, not all events that happened might be recorded correctly. Some events might be undetected (Mei *et al.*, 2019) or recorded event times may be inaccurate (Trouleau *et al.*, 2019; Shlomovich *et al.*, 2022; Guttorp *et al.*, 2023). We refer to this measurement error as data distortion, and it can occur for several reasons. For example, in earthquake catalogues, it is well known that the occurrence of large earthquakes has a masking effect which reduces the probability of subsequent earthquakes being detected for a period of time (Helmstetter *et al.*, 2006; Omi *et al.*, 2014). Alternatively, event data may simply not be available for an interval of time, such as in a terrorism data set considered by Tucker *et al.* (2019). Figure 1.1 gives a flavour of this, where observations in the shaded area are removed from our data set. This manual data distortion allows us to compare estimation procedures on missing data with the ground truth computed on the full data.

Such missing data causes serious problems for Hawkes process estimation. If the model assumes complete data and the parameters are learned using only the observed, potentially incomplete, data then parameter estimates may be severely biased. As such, a principled learning algorithm needs to consider the impact of distortion. So far, there has been limited work on learning Hawkes processes in

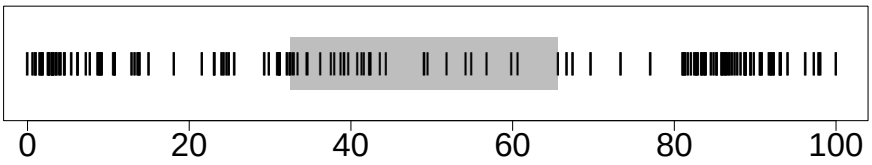


Figure 1.1: Twitter data observations. Each line indicates one event. Events inside the shaded area are removed to create the distorted data set.

the presence of distorted data. One aspect addressed in the literature is distortion in the form of gaps in the observations where no events are detected at all for a period of time (Le, 2018; Shelton *et al.*, 2018). In this context, Tucker *et al.* (2019) develop a Bayesian estimation algorithm to impute missing events, and a similar approach is proposed by Mei *et al.* (2019) using particle smoothing. Linderman *et al.* (2017) view the true generating process as a latent variable, which can be learned through sequential Monte Carlo techniques. Other examples look at specific instances of censored data (Xu *et al.*, 2017) or asynchronous data (Upadhyay *et al.*, 2018; Trouleau *et al.*, 2019).

We present a more general approach for estimating Hawkes processes in the presence of distortion, which can handle more than one distortion scenario, including the case of gaps in the observed data, and the case where there is a reduced probability of detecting events during some time period. Our approach assumes the existence of a general distortion function which specifies the type of distortion that is present. The resulting Hawkes process likelihood is computationally intractable since the self-excitation component involves triggering from the (unobserved) true event times, which must be integrated out to give the likelihood of the observed data. To solve this problem, we propose a novel estimation scheme using Approximate Bayesian Computation (ABC, Marin *et al.*, 2012; Fearnhead and Prangle, 2012) to learn the Hawkes process parameters in the presence of distortion.

In the classic Hawkes process with excitation, one event makes it *more* likely that other events occur soon after. When the occurrence of one event makes other events *less* likely, it is called *inhibition*. This can be incorporated into the Hawkes process framework but requires additional considerations. There are two main concerns. Firstly, the need for a non-negative intensity results in difficulties when computing the likelihood as integrating the intensity is not straightforward. We present exact solutions for two scenarios, as well as an approximate approach that is computationally quicker. Secondly, we found that the current criteria to assess stability when inhibition is present were unnecessarily restrictive. We propose an alternative less restrictive criterion and show that it encompasses the two existing conditions.

Hawkes processes also find application in the analysis of product sales (Pitkin *et al.*, 2018). In collaboration with a large international company, we investigate how products sell together. With a variety of similar products available,

customers need to make purchasing decisions and prioritise some products over others. For example, a consumer may choose to buy one particular good instead of a different (yet similarly designed) one. Such an effect is known as *product cannibalisation*. More formally, product cannibalisation in the marketplace is defined as the decrease in sales of one product due to the sales of a closely related product (Copulsky, 1976).

Understanding product cannibalisation is important for a variety of market participants (De Giovanni and Ramani, 2018). For example, the producers of goods can use knowledge about product cannibalisation to improve product catalogues (Child *et al.*, 1991; Desai, 2001), wholesalers can optimise for price (De Giovanni and Ramani, 2018), and retailers can make more informed decisions about which products to display on the shop floor (Kong, 2015). However, there exists "*little factual knowledge about the potential of market cannibalisation*" (Atasu *et al.*, 2010). In cooperation with our industry partner, we seek to further investigate this topic.

We choose a point process to model sales events using a variation of the Hawkes process model with inhibition, where one event makes other events less likely. This precisely represents product cannibalisation, where the purchase of one product makes it less likely that another product will be bought. Through using this particular Hawkes process we build a model with parameters that can be directly interpreted as product cannibalisation, as showcased on real data examples.

Finally, Hawkes processes have a latent variable representation, called branching structure (Daley and Vere-Jones, 2003). This is commonly used to sample events from a Hawkes process without the need to evaluate the likelihood. Conditional on the branching structure some parameters are independent, which can be exploited in a classic Gibbs sampler targeting the posterior distribution (Rasmussen, 2013; Ross, 2021). The branching structure assigns events either to immigrant or triggered events, i.e. they either come from the background process or are offsprings of other events. Despite this fundamental difference in events, all events are then treated equally in most modelling approaches. This disregard for the branching structure in the parameter structure might not be appropriate for all applications. To that end, we propose an adapted Hawkes process that allows different influence parameters for immigrant and triggered events and give an example on simulated data.

In addition, we showcase this proposed model on group chat data. This is a scenario where a handful of participants have the opportunity of equal participation (few-to-few). It sits neatly between two applications where Hawkes processes have already successfully been used; one-to-many structures such as Twitter/X (Rizoiu *et al.*, 2017) and one-to-one communication of emails (Miscouridou *et al.*, 2018). As the current research landscape into group chat data modelling is sparse (for one of the few examples see Guo *et al.*, 2019) we collect our own data set from a group chat setting to demonstrate the usefulness of our proposed model.

1.1 Contributions

This section outlines the original contributions made by this research. The work presented in this thesis contributes to the existing body of knowledge in several ways, addressing gaps and expanding the understanding of Hawkes processes. We hope that this research offers new perspectives and solutions for both the academic and practical landscape.

- **Improvement of Hawkes process estimation with missing data through Approximate Bayesian Computaiton**
  We explore the estimation of Hawkes processes when data is missing. This leads to an intractable likelihood, which rules out many common estimation techniques. To circumvent this, we propose a simulation-based inference approach using Approximate Bayesian Computation with tailor-made summary statistics that can adequately recover the posterior distribution.

- **Advances for the estimation of Hawkes processes with inhibition, including stability**
  The introduction of inhibition to Hawkes processes adds complexity to the estimation procedure, in particular regarding the non-negativity of the intensity function and its integral, which is required for the likelihood evaluation. To that end, we provide exact and approximate solutions for the integral of the intensity. In addition, we introduce a new, less restrictive condition for stability as existing conditions are unnecessarily strict under inhibition. This makes it easier for anyone, irrespective of their application, to utilise Hawkes processes with inhibition.

- **Introduction of a model for product cannibalisation application**
  We introduce a model that can adequately capture product cannibalisation. We utilise a Hawkes process model where inhibition is interpreted as product cannibalisation. For this, we propose a reparametrisation of the influence kernel that allows for a prior specification irrespective of the dimensionality of the problem. Based on this model it is possible to uncover and understand product cannibalisation.

- **Development of an adapted Hawkes model based on the branching structure, showcased on group chat data**
  Events from a Hawkes process are either immigrant or triggered events. While this underlying branching structure is commonly used for sampling and estimation, this fundamental quality of each event is not reflected in the parameter structure of a classic Hawkes process. We therefore develop a type of Hawkes process model that allows for different influences for immigrant and triggered events and provides an efficient estimation procedure. We showcase this model on group chat data, which we collected specifically for this line of work.

1.2 Thesis Outline

This thesis is structured as follows.

- We give some background information in Chapter 2, which introduces the Hawkes process both in its univariate and multivariate setting. Inference techniques, in particular Bayesian ones, and computational considerations are summarised.

- Chapter 3 addresses issues around Hawkes process estimation when data is missing. This missing data leads to an intractable likelihood, rendering common estimation techniques inapplicable or biased. To circumvent this we utilise Approximate Bayesian Computation (ABC) to estimate Hawkes process parameters. This is a likelihood-free approach that uses simulations from the data-generating process to conduct approximate Bayesian inference. We provide an adaption of ABC tailored to the particular data structure of the Hawkes process. We showcase its capabilities on simulated data and real data from retweets on Twitter/X. This body of work is available on ArXiv (Deutsch and Ross, 2020).

- In Chapter 4 we examine the Hawkes process with inhibition in more detail, as this is a necessary prerequisite for product cannibalisation modelling. We discuss a potentially negative intensity function and techniques to avoid this. In addition, we highlight problems in evaluating the likelihood, as the required integral of the intensity becomes difficult to compute. We offer an exact solution for a special case of parametrisation and an approximation for the general case. Finally, in this chapter, we introduce a new, less restrictive condition for stability as existing conditions are unnecessarily strict under inhibition.

- Chapter 5 contains our work on product cannibalisation. We examine different approaches to model product cannibalisation and offer our own model based on Hawkes processes with inhibition, which is based on the theoretical considerations of the previous chapter. To adequately capture product cannibalisation in a Bayesian framework we investigate the prior specifications and propose a reparametrisation such that priors can be chosen independently of the dimensionality of the problem. We then fit our proposed model on real data from our industry partner to uncover product cannibalisation in their product portfolio. Together with the content of the previous chapter, this piece of work is available on ArXiv (Deutsch and Ross, 2022).

- Finally, in Chapter 6 we propose an extension to the Hawkes process model that takes the latent branching structure into account. Classic Hawkes processes usually assume that every event has the same influence, irrespective of the branching structure. However, this might not be appropriate for all models. The branching structure is a popular latent variable formulation of the Hawkes process and is often used for fast data generation and parameter

estimation. We now incorporate this branching structure into the parameter structure, such that events with different latent states are allowed to have different influences. To fit this model we provide a Gibbs sampler which exploits conditional independencies between parameters. We highlight the advantages of our approach on simulated data and on a group chat data set, which we collected for this line of research. Not only is this, to the best of our knowledge, the first time Hawkes processes have been used to model group chat data, but it is also one of the first attempts altogether to investigate this particular kind of data.

- We end this thesis with a discussion of our work and highlight potential avenues for future research.

Chapter 2

Background

This chapter provides relevant background information for the thesis. It gives a summary of temporal point processes in Section 2.1. Section 2.2 then focuses on Hawkes processes, a particular type of point process used throughout all the remaining chapters. In addition, inference procedures for Hawkes processes are outlined in Section 2.3.

2.1 Temporal Point Processes

Temporal point processes are models that are used to analyse the temporal patterns of events. Whenever an event occurs, the time at which it happened is noted. This gives event times $\{t_1 \dots t_N\}$ for N events. The event times are temporally ordered such that $t_1 \leq t_2 \leq \cdots \leq t_N$ and recorded in continuous time where $t_i \in \mathbb{R}^+$ for $i = 1 \dots N$. Common examples of these types of processes are earthquakes and call centres.

Definition 2.1.1. A temporal point process on $[0, T]$ is a sequence of real, non-negative, non-decreasing random variables defined on $[0, T]$. They describe the times at which events occurred. (Laub *et al.*, 2021).

Point process models can provide an understanding of the distribution of events in time and can be used to predict when events happen in the future. They can account for seasonality, the influence of the additional covariate information, or even the impacts events have on each other. For example, events can make other events more likely, such that they cluster together, or less likely. This can be for causal reasons, or because of a data-generating process that encourages clustering. All this can be modelled with the appropriate point processes.

As their name suggests, temporal point processes contain a temporal structure where events happen one after the other. They fundamentally differ from time series models. A point process records data whenever an event occurs. In comparison, time series models take repeated measurements of an object of interest. This elementary difference warrants the usage of temporal point processes if the arrival of events is of interest.

In the following, we review the central elements of point process models. For a more detailed introduction see Cox and Isham (1980), Daley and Vere-Jones

(2003) and Laub *et al.* (2021); for a concise summary consider Laub *et al.* (2015). Ross (1995) provides a broader introduction to stochastic processes in general.

2.1.1 Counting Processes

We first examine the *counting processes* $N(t)$, which counts the number of events that happened in a time interval.

Definition 2.1.2. A counting process $N(t)$ with $t \geq 0$ is a stochastic process that satisfies all of the following conditions (Ross, 1995; Laub *et al.*, 2021):

i $N(t) \geq 0$ (it is non-negative).

ii $N(t)$ takes values in the whole numbers $\mathbb{N}^0$ (it is integer-valued).

iii If $s < t$, then $N(s) \leq N(t)$ (it is non-decreasing).

iv If $s < t$, then $N(t) - N(s)$ gives the number of events that occurred in the interval $(s, t]$.

Let us consider a point process with observed data $Y = \{y_1 \dots y_N\}$ and each $y_i = (t_i)$ for $i \dots N$. The event times are ordered such that $t_1 \leq t_2 \leq \cdots \leq t_N$ and no events happened before time 0. Furthermore, let $\mathcal{H}(t) = \{y_i : t_i < t\}$ be the set of events that happened before time t.

Then

$$N(t) := \sum_{i=1}^{N} \mathbb{1}_{\{t_i \leq t\}} \tag{2.1}$$

is a counting process. An alternative definition is

$$N(A) := \sum_{i=1}^{N} \mathbb{1}_{\{t_i \in A\}}, \tag{2.2}$$

where A is a set. When the process is plotted over time, it shows a right-continuous function with "jumps" at the event times as pictured in Figure 2.1.

2.1.2 Conditional Intensity Function

The conditional intensity function is used to define a point process. It is called *conditional*, as at time t it is conditional on everything that happened before in the process, namely $\mathcal{H}(t)$. It is defined as the expected number of events in an infinitesimal interval $[t, t + \epsilon]$, i.e.

$$\lambda(t \mid \mathcal{H}(t)) := \lim_{\epsilon \to 0^+} \frac{\mathbb{E}[N(t + \epsilon) - N(t) \mid \mathcal{H}(t)]}{\epsilon}, \tag{2.3}$$

where $\mathbb{E}[\cdot]$ denotes the expected value and $\epsilon > 0$. For simplicity, we are suppressing the conditioning on $\mathcal{H}(t)$ in the following where it is not explicitly needed. It is important to note that the intensity function cannot be negative. Hence $\lambda(t) \geq 0$ for all t needs to be ensured.

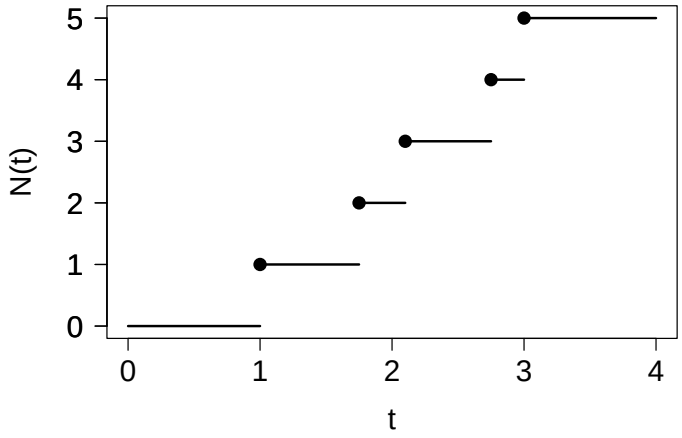


Figure 2.1: Toy example of courting process $N(t)$ for five events.

2.1.3 Poisson Processes

Let us now consider a simple point process, the inhomogeneous *Poisson process*. The inhomogeneous Poisson process is a point process with intensity function $\lambda(t)$ that has the following properties.

Definition 2.1.3. A counting process $N(t)$ is a said to be a Poisson process with rate $\lambda(\cdot) > 0$ if the following conditions hold (Ross, 1995; Daley and Vere-Jones, 2003; Laub *et al.*, 2021):

i The process has independent increments, i.e. for disjoint intervals $A_1 \dots A_k$ the random variables $N(A_1) \dots N(A_k)$ are independent.

ii For any intervals $(s, t]$, the random variable $N(t) - N(s)$ follows a Poisson distribution with mean given by the integral of the intensity function over the interval $(s, t]$:

$$\mathbb{P}\left[N(t) - N(s) = n\right] = \exp\left(-\int_s^t \lambda(u)\, du\right) \frac{\left(\int_s^t \lambda(u)\, du\right)^n}{n!},$$

for $n \in \mathbb{N}^0$ and $s < t$.

Consider a Poisson process with intensity function $\lambda(t)$ and observed events Y at times $(t_1 \dots t_N)$, where each $t_i \in [0, T]$ for $i = 1 \dots N$. Then their joint density is

$$p(Y) = \left[\prod_{i=1}^{N} \lambda(t_i)\right] \exp^{-\Lambda}, \tag{2.4}$$

where $\Lambda := \int_0^T \lambda(z)\, dz$, the integral of the intensity function over the whole observational period (Laub *et al.*, 2021)

An important property of the Poisson process is superposition. This states that if events from multiple Poisson processes are combined, they again form a Poisson process. More formally, let us assume there are K different Poisson

processes $N_1(t) \dots N_K(t)$, each with an intensity $\lambda_k(t)$ for $k = 1 \dots K$. Now consider the union of all processes

$$\mathbf{N} = \cup_{k=1}^{k} N_k. \tag{2.5}$$

It can be shown that $\mathbf{N}$ is a Poisson process with intensity $\lambda(t) = \sum_{k=1}^{K} \lambda_k(t)$ (Cox and Isham, 1980).

A special case of the Poisson process is the homogeneous Poisson process with constant intensity such that

$$\lambda(t) \equiv \lambda. \tag{2.6}$$

In that case, the interevent times as $D_i = t_i - t_{i-1}$ are independently distributed with exponential distribution with parameter λ (Cox and Isham, 1980).

2.2 Hawkes Processes

We now examine a particular kind of point process, the Hawkes process, which is utilised in all subsequent chapters of this thesis. It is often used to model events that occur in clusters or bursts. Hawkes processes are named after Alan G. Hawkes, who first defined them in his seminal paper titled "Spectra of some self-exciting and mutually exciting point processes" (Hawkes, 1971).

This section provides the conceptual and mathematical foundation of Hawkes processes for the remaining chapters of the thesis. In particular, we review Hawkes processes in one dimension (univariate) and more dimensions (multivariate). We also cover the branching structure interpretation, which is relevant, in particular, for Chapter 5 and Chapter 6.

2.2.1 Univariate Hawkes Processes

We first consider a Hawkes process in one dimension, i.e. a univariate case. This model is utilised in Chapter 3. As described above, there are a total of N observed events, with the i^{th} event recorded at time t_i. Here, $0 \leq t_i \leq T$ for all t_i where $i = 1 \dots N$.

The data Y in the univariate case is $Y = \{y_1 \dots y_N\}$ and $y_i = (t_i)$ for $i = 1 \dots N$. This notation may seem superfluous at first but is advantageous in the multivariate case below, where each event is augmented by its dimension. This then simplifies the display of certain equations afterwards.

The conditional intensity of a univariate Hawkes process at time t is

$$\lambda(t \mid \boldsymbol{\theta}, \mathcal{H}_t) = \mu(t) + \sum_{i:t_i<t} K\, g(t - t_i). \tag{2.7}$$

At time t, the intensity is conditional on $\mathcal{H}_t$, which contains all events that happened before t. For ease of notation, the dependence on $\mathcal{H}_t$ may be omitted.

Three parts contribute to the intensity function, namely the background rate $\mu(t) \geq 0$ parameterised by θ_μ, the influence magnitude $0 \leq K < 1$, and the influence kernel $g(\cdot) \geq 0$, which is parameterised by parameter θ_g. Hence the

parameters of this univariate Hawkes model are $\boldsymbol{\theta} = (\theta_\mu, K, \theta_g)$. We now review each of these three components and any restrictions on them.

Background Rate

The background rate of a Hawkes process is required to be non-negative, and often it is positive such that $\mu(t) > 0$ for all $t \geq 0$. The background rate of a Hawkes process can be a function of time, as in the definition above, capturing underlying seasonality or trends that are not driven by self-exciting behaviour. The choice of background rate is flexible and often application-specific. For example, Mohler (2013) uses a Log-Gaussian Cox process, Molkenthin *et al.* (2022) utilise a Gaussian Process to represent the background rate, and Markwick (2020) employs a Dirichlet process. When appropriate, it can also be set to a constant μ such that $\theta_\mu = \mu$.

Influence Magnitude and Kernel

At time t, each event that happened previously contributes to the intensity function. How much this contribution is, is governed by both the influence magnitude K and the influence kernel $g(\cdot)$. We take $g(\cdot) \geq 0$ to be a density such that $\int_0^\infty g(z)\, dz = 1$. This set-up permits a separate specification of the magnitude through K. Each event adds a total mass of K to the intensity; the kernel controls how this is spread out over time.

Here, it is required that $K \geq 0$, which drives the self-exciting property of the point process as each event increases the intensity function, which makes future events more likely. This will be adapted in Chapter 4 where we consider the case of $K < 0$, which is called inhibition. For now, however, we restrict ourselves to the non-negative case of excitation-only Hawkes processes. In addition, it is required that $K < 1$. This ensures that there are not infinite many events happening in finite time. Details on stability are provided in Chapter 4.

It is often assumed that the influence kernel is decreasing, which means that more recent events have more influence, but this can be adapted according to the problem at hand. The options for influence kernels are broad. In some instances, they are highly problem-specific, for example, Browning *et al.* (2021) use a histogram kernel for Covid-19 modelling. A popular choice of the influence kernel is the exponential kernel (e.g. Blundell *et al.*, 2012; Shelton *et al.*, 2018; Serafini *et al.*, 2023). It is

$$g(z) = \beta \exp^{-\beta z}, \tag{2.8}$$

with $\beta > 0$ and $z \geq 0$, where $z = t - t_i$ and $t > t_i$. For such a choice of influence kernel $\theta_g = \beta$. Hence the parameters of the Hawkes process are $\boldsymbol{\theta} = (\theta_\mu, K, \beta)$.

Now that all components of the intensity function are defined we can visualise it for a toy data set of 5 observations for a constant background rate and an exponential kernel. We also highlight the visual interpretation of the three parameters in Figure 2.2.

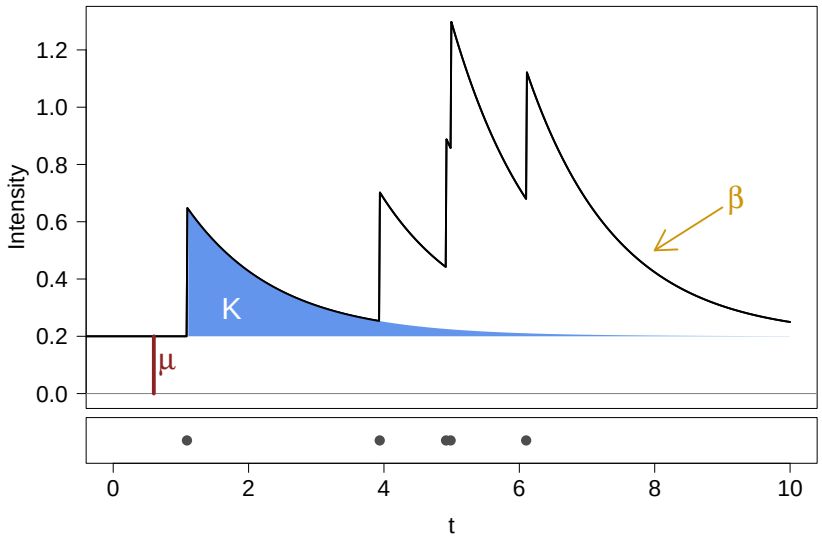


Figure 2.2: **Upper panel** shows the intensity function of a Hawkes process with $\mu = 0.2$, $K = 0.6$, and $\beta = 0.75$. Each of the parameters is schematically indicated in the intensity function. **Lower panel** plots the event times.

Branching Structure

We now examine how data from a Hawkes process can be related to an underlying branching structure. We utilise this concept in Chapter 5 to motivate a new parametrisation and subsequent prior specification, and in Chapter 6 to propose a modified Hawkes model. We consider the common branching structure interpretation when $K > 0$.

The univariate Hawkes process with intensity function from Equation 2.7 can be written as the superposition of Poisson processes such that the intensity function is a sum of independent Poisson processes. Suppose that L events $\{y_1 \dots y_L\}$ at $\{t_1 \dots t_L\}$ happened before time t. Then the intensity at time t is the sum of the background process $\mu(t)$ and L offspring processes with intensities $K\, g(t - t_i)$ for $i = 1 \dots L$, where each offspring process was *triggered* by a previous event. This gives rise to the following branching structure interpretation of a Hawkes process (Hawkes and Oakes, 1974; Daley and Vere-Jones, 2003). Data generated from an excitation-only Hawkes process consists of two types of events that come from distinct Poisson processes, where each event is generated by exactly one process (Rasmussen, 2013):

1. *Immigrant events* come from the background process with intensity $\mu(t)$.

2. *Offspring events* come from an offspring process which had been triggered by a previous event. The offspring process of event i has intensity $K\, g(t - t_i)$ for $t > t_i$. Here, each event has an average of K *direct* offsprings if $T \to \infty$.

Note that offspring events trigger offspring processes as well. This can lead to *cascades* started by an immigrant event that has offspring events, which, in turn, has offspring events etc. Figure 2.3 visualises this branching structure. This

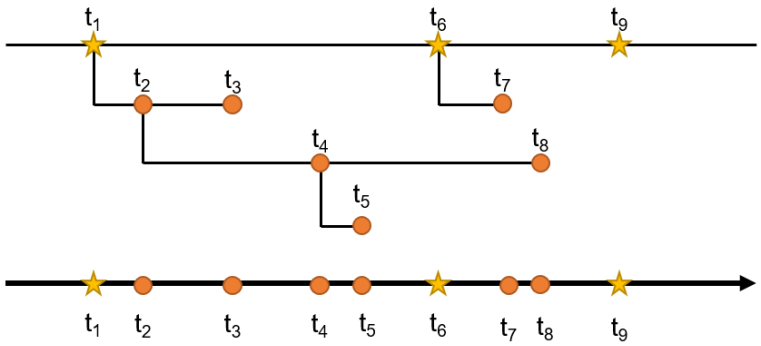


Figure 2.3: Sketch of the branching structure interpretation in one dimension. In this example, (y_1, y_6, y_9) are the immigrant events with event times (t_1, t_6, t_9), indicated by a star on the top and the summarised timeline below. The event at t_1 has two *direct* offsprings, namely at t_2 and t_3. The former has further offsprings. For example, y_1 has a total of five offsprings (number of offspring events in the cascade that started in y_1), whereas y_9 has none.

concept will be utilised in Chapter 5 to introduce a new parametrisation based on the total number of offsprings.

It is worth highlighting that the branching structure gives a direct interpretation to the parameter K, as any event has, on average, K offsprings. This interpretation holds asymptotically for $T \to \infty$. For finite T some offspring event times may be larger than T and therefore would not be included in a simulated data set. Such edge effects are common in the Hawkes literature and diminish when T is large (Daley and Vere-Jones, 2003, p. 275).

While it is not known whether an event is an immigrant or offspring for a given data set, this perspective nevertheless can be exploited for sampling, inference, and interpretation (Rasmussen, 2013; Ross, 2021). In particular, it allows for the sampling of a data set of events from a Hawkes process without having to evaluate the likelihood.

2.2.2 Multivariate Hawkes Processes

So far we have dealt with Hawkes processes where all events come from one process and influence each other equally, i.e. univariate. Now we examine the multivariate specification of the Hawkes process where events come from different processes and may have different influences on each other. Examples of multivariate Hawkes processes include the modelling of neural activity in the brain (Eichler *et al.*, 2017). Each neuron is represented by a dimension and an event is recorded whenever that neuron fires. On one hand, some neurons tend to fire together, i.e. they excite each other. On the other hand, there is a period of decreased activity for each neuron after it just fired. Such dynamics can be represented by a

multivariate Hawkes process. Similarly, in Chapter 5 each product is represented by a dimension and an event is recorded whenever the respective product is bought. The multivariate Hakwes process is also used in Chapter 4 and Chapter 6.

For the multivariate Hawkes process, data now consists of tuples for each event. More precisely, $Y = \{y_i = (t_i, d_i)\}_{i=1}^{N}$. As above, t_i gives the time at which the i^{th} event happened and $0 \leq t_i \leq T$ for all $i = 1 \dots N$. It is also assumed that the data set is ordered such that $t_1 \leq t_2 \leq \dots \leq t_N$ and events happen at the same time with probability zero. Now d_i records the dimension of the i^{th} event. We assume that each $d_i \in \{1, 2 \dots M\}$, where M, the number of dimensions, is known.

The conditional intensity function in dimension m at time t for a multivariate Hawkes process is

$$\lambda_m(t \mid \mathcal{H}_t) = \mu_m(t) + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i = j, \\ t_i < t}} K_{jm}\, g_{jm}(t - t_i). \tag{2.9}$$

Each dimension has its own background rate $\mu_m(t) \geq 0$. At time t each event contributes to the intensity function, but this contribution now comes with indices, as they reflect the influences across dimensions. This accounts for the fact that each dimension $j = 1 \dots M$ has a particular contribution to the intensity in dimension m, which is governed by $0 \leq K_{jm} < 1$ and $g_{jm}(\cdot) \geq 0$. This gives parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mu, \mathbf{K}, \boldsymbol{\theta}_g)$. As above, let us examine each component of this intensity function.

Background Rate

The background rate is now specified separately for each dimension $m = 1 \dots M$ capturing seasonality or trends. For such $\mu_1(t) \dots \mu_M(t)$ the parameters are contained in $\boldsymbol{\theta}_\mu$. As above, we require any background rate to be non-negative and often assume a positive background rate. Where appropriate a constant background rate may be assumed, in which case the parameter vector is $\boldsymbol{\theta}_\mu = \boldsymbol{\mu} = (\mu_1 \dots \mu_M)$.

Influence Magnitude and Kernel

In the univariate case, the influence magnitude was a single number. Now, for M dimensions, the influence from each dimension onto all M dimensions, including itself, is captured. More formally, K_{jm} describes what happens to the intensity in dimension m in response to an event in dimension j. We sometimes write $K_{j \to m}$ to emphasise the direction of influence, but often omit the "$\to$" in the subscript for brevity. As above, $K_{jm} \geq 0$ for $j = 1 \dots M$ and $m = 1 \dots M$ is assumed. Any negative K_{jm} would imply inhibition. This subtle extension requires additional considerations, which we investigate in Chapter 4. For now, we limit ourselves to the excitation-only case. As above, we assume that $g_{jm}(\cdot) \geq 0$ and $\int_0^\infty g_{jm}(z)\, dz = 1$ for all $j, m = 1 \dots M$.

The parameters are written as matrix $\mathbf{K} = \{K_{jm}\}$ where $j, m = 1 \dots M$. Note that no symmetry or any other structure in $\mathbf{K}$ is enforced. To ensure stability,

the spectral radius (i.e. the largest absolute eigenvalue) of $\mathbf{K}$ is required to be less than 1. Stability is further investigated in Chapter 4.

As just seen in the influence magnitude, the influence kernel is now also equipped with indices indicating the direction of influence from one dimension to another. Hence, $g_{jm}(\cdot)$ governs the decay in influence from an event that happened in dimension j onto dimension m. The parameters used in the influence kernel are contained in $\boldsymbol{\theta}_g$. The exponential decay remains a popular choice in the multivariate Hawkes where

$$g_{jm}(z) = \beta_{jm} \exp^{-\beta_{jm} z}, \tag{2.10}$$

for $z > 0$ with each $\beta_{jm} > 0$ and $z > 0$. In that case, the parameters are arranged in a matrix such that $\boldsymbol{\beta} = \{\beta_{jm}\}_{j,m=1\ldots M}$.

Branching Structure

While Section 2.2.1 focuses on the univariate case, the extension to multivariate Hawkes processes is straightforward (see, for example, Embrechts *et al.*, 2011). In the multivariate case (with all entries of $\mathbf{K}$ being non-negative) an event in dimension j triggers offspring processes in all dimensions $m = 1 \ldots M$, including where $j = m$. Here, each event in dimension j has on average K_{jm} offsprings in dimension m. In turn, these offspring events can trigger offspring processes in all dimensions. Hence, cascades can potentially lead through different dimensions, which all contribute to the total number of offsprings of the immigrant event that started the cascade, which is discussed in Chapter 5. Chapter 6 uses the branching structure in the multivariate case to motivate two novel models.

2.3 Inference

This section reviews the likelihood functions for both univariate and multivariate Hawkes processes, which are used throughout this thesis. In addition, we highlight common inference techniques for Hawkes processes, in particular through generating samples from the posterior distribution from a Bayesian point of view.

Note that throughout this section we do not take into account events that happened before the observational period at $t < 0$. For some applications, such as earthquakes, it may be important to incorporate those events, as they can have offspring that were observed at $t > 0$. When past events are unknown, the resulting edge effects can be rectified through, for example, "perfect" sampling (Møller and Rasmussen, 2005).

2.3.1 Likelihood

This section covers the likelihood functions for data Y with events in $[0, T]$, both in the univariate and multivariate setting.

Univariate

As for any point process, the likelihood of a univariate Hawkes process with intensity $\lambda(\cdot)$ for data Y is

$$p(Y \mid \boldsymbol{\theta}) = \left[\prod_{i=1}^{N} \lambda(t_i \mid \boldsymbol{\theta})\right] \exp^{-\Lambda}, \tag{2.11}$$

where $\Lambda := \int_0^T \lambda(t \mid \boldsymbol{\theta})\, dt$ (Daley and Vere-Jones, 2003). For a kernel $g(\cdot)$ with corresponding integral $G(\cdot)$ the integral of the intensity can be expressed as

$$\Lambda = \int_0^T \lambda(t \mid \boldsymbol{\theta})\, dt = \int_0^T \mu(t)\, dt + K \sum_{j=1}^{N} G(T - t_j), \tag{2.12}$$

which simplifies the likelihood evaluation when $G(\cdot)$ has a closed-form solution. For a constant background rate $\mu(t) = \mu$ the integral $\int_0^T \mu(t)\, dt$ is just μT.

For a univariate Hawkes process with an exponential influence kernel

$$\Lambda = \int_0^T \mu(t)\, dt + K \sum_{j=1}^{N} \left(1 - \exp^{-\beta(T - t_j)}\right) \tag{2.13}$$

gives a convenient way to rewrite the integral of the intensity function (Laub *et al.*, 2021).

Multivariate

For a multivariate Hawkes process the parameters are $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mu, \mathbf{K}, \boldsymbol{\theta}_g)$. The likelihood for observed data Y is analogous to the univariate case. We define $Y_m := \{y_i : d_i = m\}$, i.e. the observations in dimension m for $m = 1 \ldots M$. For any multivariate point process, the likelihood for data Y is

$$p(Y \mid \boldsymbol{\theta}) = \prod_{m=1}^{M} p(Y_m \mid \boldsymbol{\theta}) \tag{2.14}$$

$$= \prod_{m=1}^{M} \left[\prod_{i : y_i \in Y_m} \lambda_m(t_i \mid \boldsymbol{\theta})\right] \exp\left(-\Lambda_m\right), \tag{2.15}$$

where $\Lambda_m = \int_0^T \lambda_m(t \mid \boldsymbol{\theta})\, dt$ (Daley and Vere-Jones, 2003). To compute each Λ_m we can make use of the univariate simplification described in Equation (2.13).

2.3.2 Bayesian Estimation

For Hawkes process inference, both frequentist and Bayesian estimation and inference methods have been widely adopted. For example, Veen and Schoenberg (2008) make use of an expectation maximisation algorithm in the frequentist domain, Lemonnier and Vayatis (2014) use a non-parametric approach, and Zuo

et al. (2020) employ a transformer architecture. Rasmussen (2013) provides a seminal paper on Bayesian estimation of Hawkes processes, including some computationally efficient sampling techniques based on the branching structure. Ross (2021) extends this line of work for the ETAS model. Serafini *et al.* (2023) give an estimation procedure using an integrated nested Laplace approximation.

Most of this thesis makes use of the Bayesian paradigm. This uses both the likelihood $L(Y \mid \theta)$ for data Y under parameter $\theta \in \Theta$, as well as a prior distribution $\pi(\theta)$ for parameter θ. Together they specify the posterior distribution $\pi(\theta \mid Y)$ as

$$\pi(\theta \mid Y) = \frac{\pi(\theta)L(Y \mid \theta)}{\int_\Theta \pi(\theta')L(Y \mid \theta')\, d\theta'}. \tag{2.16}$$

The fact that $\pi(\theta \mid Y)$ is a distribution facilitates a variety of analyses. Inference is not confined to a single point estimate with error bars but can explore the whole distribution, including its quantiles, shape, and other characteristics. However, in many cases, it is difficult or impossible to find a closed-form solution for the posterior distribution. Instead, Monte Carlo methods are employed, where random samples from a distribution are used to approximate it (Robert and Casella, 2004). This is particularly relevant in the Bayesian framework, as it is often possible to generate samples from the posterior distribution, even when it does not have a closed form.

This need to generate samples from distributions opens up a whole field of research focusing on sampling techniques. An important building block is the Metropolis-Hastings (MH) algorithm, which produces samples from a target distribution $p(\cdot)$. In our case, we take $p(\theta)$ to be the posterior distribution, but MH can be used to sample from many distributions of interest. We now provide a sketch of the MH procedure, as this is utilised explicitly in Chapter 3 and Chapter 6, and implicitly in Chapter 5.

2.3.3 Metropolis-Hastings

The goal of MH is to produce samples from any distribution $p(\cdot)$. To sample from $p(\cdot)$ via MH $p(\theta)$ needs to be able to be evaluated for all $\theta \in \Theta$. In addition a proposal distribution $q(\cdot \mid \cdot)$ is specified, which can be evaluated and sampled easily. MH also requires an initial value θ_0 to start the iterative sampling procedure.

At step k of the sampling value $\theta^* \sim q(\cdot \mid \theta^{(k-1)})$ is proposed. With probability

$$\min\left(\frac{p(\theta^*)}{p(\theta^{(k-1)})}\frac{q(\theta^{(k-1)} \mid \theta^*)}{q(\theta^* \mid \theta^{(k-1)})}, 1\right) \tag{2.17}$$

we set $\theta^{(k)} = \theta^*$; otherwise we set $\theta^{(k)} = \theta^{(k-1)}$. It can be shown that this set-up targets $p(\cdot)$, the distribution of interest.

This is repeated until some predefined stop criterion has been reached. To minimise the dependence on the starting value $\theta^{(0)}$ the first $30 - 50\%$ of the samples are usually discarded as burn-in to remove dependence on the starting value. The remaining samples may be thinned, often such that only every 10^{th} sample is retained. This can be due to storage reasons but is not strictly necessary

for the performance of the sampler. These MCMC samples then come from the desired distribution $p(\cdot)$.

As this description only serves as an outline of MH sampling, a detailed account of sampling diagnostics and many other intricacies lies beyond the scope of this work. Please see Gelman *et al.* (in particular Chapter 11, 2013) or Robert (2016) for more comprehensive discussions.

2.3.4 Stan

While the above description gives the blueprint to sample from any distribution, it can be difficult to obtain high-quality samples via MH when target distributions are "difficult", e.g. multimodal or high dimensional. For that reason, there are many algorithms built on top of the MH framework, such as Sequential Monte Carlo (Del Moral *et al.*, 2012) or Hamiltonian Monte Carlo (Betancourt, 2017) that improve on the vanilla MH procedure.

To make these usable for researchers there are a variety of implementations readily available. In this thesis, we make particular use of Stan through `rstan`, which provides the relevant R interface (Stan Development Team, 2019). Stan utilises a No-U-Turn sampler (a variation of Hamiltonian Monte Carlo) to obtain samples from a posterior distribution. This library specialises in Bayesian statistics, as it permits the user to define priors and likelihood in a syntax similar to R that is then translated into C++. This enables fast posterior sampling from Hawkes processes in a Bayesian paradigm such that we can conduct the relevant posterior inference. We make use of Stan in R in Chapter 3 and Chapter 5.

Chapter 3

ABC for Hawkes Processes with Missing Event Times

Most applications of the Hawkes process assume that the entire history of events has been accurately and completely observed. However, many real-world applications of point processes suffer from missing data where some events are undetected (Mei *et al.*, 2019), or from noisy data where the recorded event times are inaccurate (Trouleau *et al.*, 2019; Shlomovich *et al.*, 2022; Guttorp *et al.*, 2023). We refer to this as data distortion, and it can occur for several reasons. For example, in earthquake catalogues, it is well known that the occurrence of large earthquakes has a masking effect which reduces the probability of subsequent earthquakes being detected for a period of time (Helmstetter *et al.*, 2006; Omi *et al.*, 2014). Alternatively, event data may simply not be available for an interval of time, such as in a terrorism data set considered by Tucker *et al.* (2019). This also parallels the initialisation problem, where events before $t = 0$ did happen, but are not recorded (Møller and Rasmussen, 2005). Furthermore, for point processes in ecology imperfect detection is a common issue, for example when counting population via satellite images (Williams *et al.*, 2017). These models can also suffer from presence-only data, where it is only recorded where a species was present, but not when it was not present (Renner *et al.*, 2015).

Distorted (e.g. missing or noisy) data cause serious problems for Hawkes processes. If the model assumes complete, correct data and the parameters are learned using only the observed, potentially incomplete, data then the estimation of $\boldsymbol{\theta}$ may be severely biased. As such, a principled approach needs to consider the impact of distortion. Some research on Hawkes process inference in the presence of distorted data has focused on gaps in the observations where no events are detected at all for a period of time (Le, 2018; Shelton *et al.*, 2018). In this context, Tucker *et al.* (2019) develop a Bayesian estimation algorithm which uses MCMC to impute missing events, and a similar approach is proposed by Mei *et al.* (2019) using particle smoothing. Linderman *et al.* (2017) view the true generating process as latent variables, which can be learned through sequential Monte Carlo techniques. Other examples look at specific instances of censored data (Xu *et al.*, 2017), censored data in junction with a known number of missing event times (Kong *et al.*, 2023), or asynchronous data (Upadhyay *et al.*, 2018; Trouleau *et al.*, 2019).

In this chapter, we present a more general approach for estimating Hawkes processes in the presence of distortion, which can handle more than one distortion scenario, including the case of gaps in the observed data, and the case where there is a reduced probability of detecting events during some time period. Our approach assumes the existence of a general distortion function $h(\cdot)$ which specifies the type of distortion that is present. The resulting Hawkes process likelihood is computationally intractable since the self-excitation component involves triggering from the (unobserved) true event times, which must be integrated out to give the likelihood of the observed data. To solve this problem, we propose a novel estimation scheme using Approximate Bayesian Computation (ABC, Marin *et al.*, 2012) to learn the Hawkes intensity in the presence of distortion.

The chapter is organised as follows. In Section 3.1 we introduce the Hawkes process and the distorted data setting. Section 3.2 summarises ABC and introduces the ABC-Hawkes algorithm for parameter inference in the presence of distorted data. In Section 3.3 we consider different types of distortion, namely gaps in the data and missing observations, and discuss some theoretical results relating to identifiability. Section 3.4 investigates the performance of our ABC algorithm using both simulated and real data with different types of distortion, both when the distortion parameters are known and unknown.

3.1 Problem Overview

In this section, we propose the implementation of distortion functions. We highlight the problem arising from missing or noisy events, which causes the Hawkes likelihood function to become computationally intractable.

As defined in Chapter 2, a univariate Hawkes process with a constant background rate and exponential decay kernel is used. For data $Y_{\text{complete}} = \{y_1 \dots y_L\}$ and $y_i = (t_i)$ for $i = 1 \dots L$ with $t_i \in [0, T]$ the conditional intensity is

$$\lambda(t \mid \mathcal{H}_t, \boldsymbol{\theta}) = \mu + \sum_{\substack{i: t_i < t, \\ y_i \in Y_{\text{complete}}}} K\,\beta\,\exp(-\beta(t - t_i)). \tag{3.1}$$

Hence, the parameters in this set-up are $\boldsymbol{\theta} = (\mu, K, \beta)$. However, our method can be adapted for other specifications.

3.1.1 Distorted Data

In many applications, the observed data will be distorted, for example containing missing or inaccurately recorded events. This is often due to data collection issues which result in some events being undetected or observed with error. This can also be described as measurement error, for example, synchronisation errors (Trouleau *et al.*, 2019) or binned data (Shlomovich *et al.*, 2022).

In the case of missing data, let $h(\cdot)$ be a distortion function that specifies the probability that an event occurring at time t will be successfully detected, and hence that it will be present in the observed data Y (Ringdal, 1975; Omi *et al.*, 2014). In some cases, the distortion function may depend on an (un-

known) parameter which we write as $h(\cdot \mid \xi)$. For now, we will suppress this in our notation and assume that all parameters of the distortion function are fully known, although we return to the unknown ξ case later.

The observed data can be viewed as having arisen from the following generative process: First, a set of event times $(t_1, \ldots, t_L)$ are generated from a Hawkes process with some intensity function $\lambda(\cdot)$. Then, for each event t_l for $l = 1, \ldots, L$, define an indicator variable D_l such that

$$\mathbb{P}[D_l = 1] = h(t_l), \tag{3.2}$$
$$\mathbb{P}[D_l = 0] = 1 - h(t_l). \tag{3.3}$$

Here, D_l indicates whether event y_l was observed. The observed data is then the collection of events for which $D_l = 1$ so that $Y = \{y_l \mid l : D_l = 1\}$. This specification is highly flexible and also covers scenarios where events are not missing for a certain period of time, but are instead missing with a (possibly time-dependent) non-zero probability, as in the earthquake scenario. This also encompasses the initialisation problem, where the history of events before the observational period is unknown (Møller and Rasmussen, 2005).

When data is potentially missing, it is very challenging to learn the parameters of the Hawkes process. If the Hawkes process did not have a self-exciting component, then the likelihood function in the presence of missing data would be obtained by assuming that the observed data came from a modified point process with intensity $f(t) = \lambda(t)h(t)$, i.e. the product of the intensity function and the distortion function. The parameters of $f(\cdot)$ could then be learned using a standard method such as maximum likelihood or Bayesian inference. However, when working with Hawkes processes, the situation is substantially more complex. The main issue is that undetected events will still contribute to the intensity, i.e. the summation in Equation 3.1 needs to be over both the observed and unobserved events. The resulting likelihood function hence depends on both the set of observed events Y and the set of unobserved events which we denote by Y_u. This requires "integrating out" the unobserved events, which gives the likelihood function

$$p(Y \mid \boldsymbol{\theta}) \propto \int p(Y, Y_u \mid \boldsymbol{\theta}) \prod_{t_i \in Y} h(t_i) \prod_{t_j \in Y_u} (1 - h(t_j))\, dY_u, \tag{3.4}$$

where $Y_u = \{t_k \mid k : D_k = 0\}$ denotes the set of unobserved (=missing) events. Due to the integral over the unknown number of missing events, this likelihood function is intractable and cannot be evaluated, which renders many inference approaches unusable. One approach is to treat the missing events Y_u as latent variables, which can be imputed, as proposed by Tucker *et al.* (2019). We now propose an alternative, novel scheme for Hawkes process inference with distorted data based on Approximate Bayesian Computation which we refer to as ABC-Hawkes.

3.2 Approximate Bayesian Computation

ABC is a widely studied approach to Bayesian inference in models with intractable likelihood functions (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Marin *et al.*, 2012). We will now review the general ABC framework and then present a version of ABC for sampling from the posterior distribution of the Hawkes process parameters in the presence of distorted data.

Bayesian inference for a parameter vector $\theta \in \Theta$ assumes the existence of a prior $\pi(\theta)$ and a likelihood function $L(Y \mid \theta)$ for data $Y \in \mathcal{Y}$, with parameter inference based on the resulting posterior distribution $\pi(\theta \mid Y)$. Traditional tools for posterior inference using Markov Chain Monte Carlo (MCMC) require the evaluation of the likelihood function and therefore cannot be applied when the likelihood is intractable. In this case, however, it may still be possible to generate samples $Y^{(j)}$ from the model in a way that does not require a likelihood evaluation. ABC is an approach to parameter and posterior density estimation which only uses such generated samples, without the need to evaluate the likelihood (Beaumont *et al.*, 2002).

The core idea of ABC is as follows: It is assumed that the observed data Y has been generated by some (unknown) value of θ. For a proposed value of $\theta^{(j)}$, a pseudo-data $Y^{(j)}$ is generated from $L(Y \mid \theta^{(j)})$ in a way which does not involve evaluating the likelihood function. If $\theta^{(j)}$ is close to the real θ, then $Y^{(j)}$ is also expected to be "close" to the real (observed) data Y, as measured by a similarity function. Hence, one can accept/reject parameter proposals $\theta^{(j)}$ based only on the similarity between Y and $Y^{(j)}$.

This algorithm crucially depends on how similarity is measured between data sets Y and $Y^{(j)}$. Typically, low-dimensional summary statistics $S(\cdot)$ of the data are chosen and then compared based on some distance metric $\mathcal{D}(\cdot, \cdot)$ (Fearnhead and Prangle, 2012). A proposed parameter $\theta^{(j)}$ is then accepted if this distance between data sets is less than a chosen threshold ϵ.

A direct implementation of the ABC procedure is based upon rejection sampling, where $\theta^{(j)}$ is proposed from the prior distribution (Pritchard *et al.*, 1999). However this can be inefficient, so instead ABC-MCMC methods with Metropolis-Hastings proposals can be used. These can lead to a higher acceptance rate (Beaumont *et al.*, 2002).

Generally, ABC procedures will not target the true posterior $\pi(\theta \mid Y)$, but instead target the ABC posterior $\pi_{ABC}(\theta \mid \mathcal{D}(S(Y), S(Y^{(j)}) < \epsilon)$. However, if the statistics $S(\cdot)$ are chosen to be the sufficient statistics for the model parameters and $\epsilon \to 0$, then $\pi_{ABC}(\theta \mid \mathcal{D}(S(Y), S(Y^{(j)})) < \epsilon) \to \pi(\theta \mid Y)$ (Marin *et al.*, 2012). In most realistic applications it is not possible to choose a low-dimensional set of summary statistics which is sufficient for the parameter vector since only the limited class of distributions that lie in the exponential family admit a finite-dimensional set of sufficient statistics (Brown, 1986).

Therefore, finding suitable summary statistics and appropriate distance metrics for a particular model is non-trivial, and is a vital part of designing effective ABC algorithms (Marin *et al.*, 2012). Some authors construct summary statistics which are carefully tailored to their application (Aryal and Jones, 2020), while others present a semi-automatic way to constructing summary statistics (Fearn-

head and Prangle, 2012). Bernton *et al.* (2019) develop a general approach, that uses the Wasserstein Distance between empirical distributions of the observed and synthetic data set, hence eliminating the need for summary statistics altogether.

3.2.1 ABC-Hawkes Algorithm

Our core idea is to use ABC to perform inference for the Hawkes process with distorted data, which circumvents the intractability of the likelihood function. This is made possible by the fact that simulating data from the distorted generative model is straightforward, and can be done by first simulating data from the Hawkes process with intensity function $\lambda(\cdot)$ which represents the true (unobserved) data and then distorting these events based on the distortion function $h(\cdot)$ as discussed in Section 3.1.1, which gives the observed data Y. The simulation from $\lambda(\cdot)$ can be carried out using a standard simulation algorithm for the Hawkes process such as the thinning procedure of Ogata (1981). The distortion of the data is then performed by applying $h(\cdot)$ to each simulated data point, which introduces missingness. For example in the case of gaps, this would consist of deleting the simulated observations which lie within the gap region. The resulting observations are hence a realisation of the Hawkes process with intensity function $\lambda(\cdot)$ that has been distorted through $h(\cdot)$. This means it is possible to sample from the data-generating process without having to evaluate the likelihood.

For ABC-Hawkes we use a variant of the ABC-MCMC algorithm, as shown in Algorithm 1. This is an extension of the usual Metropolis-Hastings algorithm (see Chapter 2) which essentially replaces the intractable likelihood function with an estimate based on the simulated data and can be shown to converge correctly to the ABC posterior $\pi_{ABC}(\theta \mid \mathcal{D}(S(Y), S(Y^{(j)})) < \epsilon)$ (Marjoram *et al.*, 2003).

Algorithm 1 ABC-Hawkes (ξ known)

Input: observed data Y where data is distorted according to a distortion function $h(\cdot \mid \xi)$ with known ξ, prior $\pi(\cdot)$, desired number of posterior samples J, function to compute the P summary statistics $S_1(\cdot), \ldots, S_P(\cdot)$, P separate threshold levels $\epsilon_p > 0$, and a Metropolis-Hastings transition kernel $q(\cdot \mid \cdot)$

Initialise $\theta^{(0)}$
for $j = 1$ **to** J **do**
  $\theta^* \sim q(\cdot \mid \theta^{(j-1)})$
  $Z^* \sim p(\cdot \mid \theta^*)$, where $p(\cdot)$ simulates from a Hawkes process
  Sample Y^* such that each event in Z^* is included in Y^* with probabilities $h(Z^* \mid \xi)$
  if $\mid S_p(Y^*) - S_p(Y) \mid < \epsilon_p$ for all $p = 1, \ldots, P$ **then**
    With probability $\min\left\{1, \frac{q(\theta^{(j-1)} \mid \theta^*)\pi(\theta^*)}{q(\theta^* \mid \theta^{(j-1)})\pi(\theta^{(j-1)})}\right\}$ set $\theta^{(j)} = \theta^*$
  else
    Set $\theta^{(j)} = \theta^{(j-1)}$
  end if
end for
Output: $(\theta^{(1)}, \ldots, \theta^{(J)})$

The choices of summary statistics $S(\cdot)$ and corresponding threshold ϵ are crucial for the application of ABC (Marin *et al.*, 2012). While the ABC literature offers a wealth of theory on summary statistics, their actual construction is less prominent and tends to be highly application-dependent. One new development that has the potential to be exploited is the work by Cavaliere *et al.* (2023). They propose a way to construct bootstrap samples for Hawkes processes, which outputs pseudo-samples of event series in typical bootstrap fashion. This could unlock ABC-based methods that rely on multiple sets of observations. For example, Pudlo *et al.* (2016) use a random forest to construct summary statistics. Gutmann *et al.* (2018); Li *et al.* (2018) utilise a classifier or reinforcement learning to judge the similarity between data sets, which eliminates the need for explicit summary statistics altogether. It remains to be seen how these methods perform on the complicated non-i.i.d. data structure of the Hawkes process.

We chose to work with a small number of summary statistics for ABC-Hawkes. While there has been some previous literature on ABC for Hawkes processes outside of the distorted data setting (Ertekin *et al.*, 2015; Shirota and Gelfand, 2017, the latter spatial) the presented summary statistics did not extend well to distorted data. Instead, through extensive simulations, we have identified a set of summary statistics which we have empirically found to accurately capture the posterior distribution of the Hawkes process parameters, and in Section 3.4 we provide evidence to support this.

For the summary statistics, we make use of the interevent times, defined as

$$\begin{aligned} &\Delta = (\Delta_1, \Delta_2 \dots \Delta_{N-1}), \\ &\Delta_i = t_{i+1} - t_i \text{ for } i = 1 \dots N-1. \end{aligned} \tag{3.5}$$

For ABC-Hawkes, the resulting summary statistics $S(Y) = (S_1(Y), \dots, S_7(Y))$ we use are:

- $S_1(Y)$: The logarithm of the number of observed events in the process. This is highly informative for the μ and K parameters, since μ controls the number of background events, while K governs how much total triggering is associated with each event.

- $S_2(Y)$: The median of the interevent times Δ divided by their mean $\mathbb{E}[\Delta]$. This is highly informative for the parameters of the self-excitation kernel (e.g. β, for an exponential kernel).

- $S_3(Y) \dots S_5(Y)$: Ripley's K statistic (Ripley, 1977, not to be confused with the Hawkes parameter K) for a window size $w \in \{1, 2, 4\}$, respectively. This means that for each event y_i for $i = 1 \dots N$ we count how many events happen in the interval $(t_i, t_i + w]$. We then calculate the arithmetic mean of these counts. We do not use a border correction. We found that our choice of window sizes worked well when events happen on the order of magnitude of one event per unit of time, but the window size can be adjusted according to the scale of the data set. These summary statistics capture the degree of clustering in the event sequence and are hence informative for the parameter K and the parameters of the excitation kernel.

- $S_6(Y)$: The average of the Δ_i that lie above their 90%-quantile $\mathbb{E}[\Delta_i \mid \Delta_i > q_{90}]$. This gives a notion of how large the largest interevent times are.

- $S_7(Y)$: The average of the Δ_i that lie below their median $\mathbb{E}[\Delta_i \mid \Delta_i < q_{50}]$. This broadly measures how large the "not so large" interevent times are.

This set of summary statistics was constructed after extensive simulation studies. We systematically constructed parameter combinations that could portray a variety of scenarios (e.g. low/medium/high K with a fast/slowly decreasing kernel) and sampled event series for each. To start, we kept the events undistorted. Our first ABC posterior samples were based on a small set of summary statistics. These were then compared against the ground truth, which we could obtain from Stan as we were working with complete data.

Then we started adding additional summary statistics in different constellations and checked which ABC posteriors matched the truth best according to posterior mean, posterior standard deviation, and a visual inspection of the posterior shape. Once a candidate set of summary statistics was identified, we performed a backwards selection removing each in turn to determine whether performance deteriorated across different data sets. This allowed us to identify summary statistics that were closely correlated and to choose which ones of those to keep.

Up until that point, all work was done on complete data sets, as these permitted a comparison to the ground truth. We then took the set of summary statistics we had identified on complete data and put it through similar rigorous tests using a variety of distortion scenarios. This allowed us to determine what impact distortion had on the summary statistics and subsequently the performance of the ABC. The references we used for this were the complete-data posterior distributions obtained via Stan from the undistorted data sets. As above, we investigated performance along posterior mean, standard deviation and shape compared to the reference when swapping out individual summary statistics for closely correlated alternatives. This procedure results in a set of summary statistics that performed sufficiently well across many simulations, i.e. the seven statistics outlined above.

We also put considerable effort into an alternative approach, namely semiautomatic ABC (Fearnhead and Prangle, 2012). When we submitted the content of this chapter to a journal a reviewer pointed out that the semiautomatic approach to ABC summary statistic selection, where a large number of summary statistics is linearly weighted, might be more effective than using the "handcrafted" summary statistics above. However, despite our best efforts, the results when using the semiautomatic method were inferior, in particular for certain distortion scenarios. We believe that this is because summary statistics relate to the parameter in an intricate, non-linear fashion, which cannot be adequately captured by the linear combination suggested by the semiautomatic approach. Further details of this can be found in Appendix A.2. Hence, we instead propose to use the seven summary statistics above.

When using multiple summary statistics, careful consideration must be made when combining them. We choose to work with separate thresholds ϵ_p for each of the summary statistics. We set each ϵ_p to be a fraction of the empirical standard

deviation of the summary statistic based on a pilot run of the simulation, such that $0.01 - 0.1\%$ of the proposals are accepted (Vihola and Franks, 2020). We hence accept a proposed value $\theta^{(j)}$ only if $\mathcal{D}(S_p(Y), S_p(Y^{(j)})) < \epsilon_p$ for all p, using the absolute value function (L_1 norm) as the distance metric $\mathcal{D}(\cdot, \cdot)$. We found that the L_1 distance (separately for each of the seven summary statistics) provides the best performance for ABC-Hawkes. We consider it beneficial that each of our seven summary statistics needs to lie beneath a threshold, rather than them being summed up in an L_2 norm where their average needs to be beneath a certain value.

Finally, in the case where the distortion function $h(\cdot)$ is parameterised by unknown distortion parameters ξ, these can also be sampled from their posterior distribution, as showcased in Section 3.4.3. At each iteration of the MCMC one of the following is selected at random to be updated (any remaining are kept at the current value): (1) Hawkes parameter $\boldsymbol{\theta}$, (2) distortion parameter ξ, or (3) both $\boldsymbol{\theta}$ and ξ. The transition kernel is Gaussian of appropriate dimension with standard deviations specified below. The same summary statistics $S(Y) = (S_1(Y), \ldots, S_7(Y))$ are used for all update steps, regardless of which parameters are newly proposed.

3.3 Specific Types of Distortion

We now examine different types of data distortion expressed through the distortion function $h(\cdot \mid \xi)$. In some situations, the parameter ξ will be known due to knowledge of the distortion process (Tucker *et al.*, 2019). In other scenarios, they may be learned from the data. In the following, we examine two specific distortion scenarios in more detail, namely gap and constant deletion.

3.3.1 Gap in the Data

One of the scenarios we consider is when no observations are recorded for a given period of time, corresponding to a "gap" in the data. This may be due to a physical fault in the device used for detection, or by post-hoc data distortion. Here, we consider a single gap in the data, but this work could be extended to handle multiple gaps. In our notation, a gap implies the following distortion function

$$h(t \mid \xi) = \begin{cases} 0, & \text{if } t > g_{start} \text{ and } t < g_{end} \\ 1, & \text{otherwise} \end{cases} \tag{3.6}$$

with distortion parameter $\xi = (g_{start}, g_{end})$, which indicates the start and end of the gap, respectively. It is reasonable to assume that ξ is known, for example in the case where physical index cards recording terrorism events have gone missing (Tucker *et al.*, 2019) and the gap is clearly identified by the absence of index cards.

3.3.2 Constant Deletion

Another distortion mechanism sees each event being recorded with a constant probability

$$h(t \mid \xi) = a, \tag{3.7}$$

with parameter $\xi = a$, which is equivalent to a "constant deletion" with probability $(1 - a)$. This distortion represents, for example, the inherent failure probability of recording devices missing some events. In such cases, a could be known as it quantifies the reliability of a device, which may be provided by the manufacturer. There are other scenarios where a will not be known (e.g. Shelton *et al.*, 2018) and must be estimated from the data. This brings challenges of identifiability, as evident in our experiments in Section 3.4.3. To highlight this difficulty, recall that a Poisson process with constant deletion would not be identifiable. Hence, problems of identifiability arise when the Hawkes process is "close" to a Poisson process.

When devices fail to detect smaller earthquakes for an amount of time after a mainshock (Helmstetter *et al.*, 2006), the distortion mechanism may also depend on $\mathcal{H}_t$, the history of events. The distortion mechanism could then have change points, where the probability of detection is not constant across the whole period, but only within intervals, which can depend on the observed data. This would be a substantial expansion and lies outside the scope of this work.

3.4 Experimental Results

We now present experimental results to show the performance of the ABC-Hawkes algorithm. First, we will give evidence that the set of summary statistics we presented above captures most of the information in the parameter posterior distribution for the Hawkes process when distorting is not present.

Next, we investigate how accurately the parameters are estimated in the presence of distortion. For this, we manually insert data distortion into a simulated event sequence where we have access to the true event times. This allows us to compare the posterior distribution estimated by ABC-Hawkes on the distorted data to the complete-data posterior which would have been obtained if we had access to the undistorted data. Note that this is not the "true" posterior under the full data-generating process (Hawkes process + distortion), but instead the complete-data posterior we would obtain without distortion. Nevertheless, this is a relevant point of reference for our comparisons.

We examine three distortion scenarios: (1) gap deletion with known start and end of the gap, (2) constant deletion with known a, and (3) constant deletion with unknown a. Where available, results are compared to other approaches. Additional examples on more data sets can be found in Appendix A.1.

3.4.1 No-distortion Setting

We first confirm that the above ABC summary statistics accurately allow the posterior distribution to be estimated in a standard Hawkes process without

distortion. We then compare our approach to that of Ertekin *et al.* (2015). The latter is the only existing example we could find of applying ABC to Hawkes processes, although they do not consider the data distortion setting.

Recover Posterior Distribution

To investigate the capabilities of our algorithm to recover the complete-data posterior distribution without distortion we generate a data set and compare the posterior estimates from ABC-Hawkes to the ground truth from Stan (Stan Development Team, 2019). We use the following priors:

$$\begin{aligned} \mu &\sim \mathcal{U}(0.05, 0.85), \\ K &\sim \mathcal{U}(0, 0.9), \\ \beta &\sim \mathcal{U}(0.1, 1). \end{aligned}$$

The restriction that $K < 1$ is standard, and ensures that data sampled from the Hawkes process contain a finite number of events with probability 1. As shown in Figure 3.1 and Appendix A.1.1 ABC-Hawkes can approximate the posterior distributions, both in location and shape, in this simulation study. We note that the overestimation of the posterior variance is an often observed issue in ABC stemming from a necessary non-zero choice of the ϵ_p threshold (Li and Fearnhead, 2017). Estimation accuracy and speed could be further improved upon, for example, using approaches suggested by Fukumizu *et al.* (2013) utilising reproducing kernel Hilbert spaces.

Comparison to Ertekin et al.

The only example of ABC in the Hawkes literature we could find comes from Ertekin *et al.* (2015), who use it to estimate a model for failures in New York City's underground power grid, which is accessed via manholes. They use a point process that has both an excitation (failures, such as fire, can cascade) and an inhibition (repairs lead to fewer failures) component, as well as a term to deal with zero-inflation (failures are, fortunately, quite rare). While it is not completely clear why Ertekin *et al.* (2015) choose an ABC approach in the first place, we assume it is to avoid problems induced by inhibition, such as the need to integrate the intensity (which we describe in Chapter 4).

Ertekin *et al.* (2015) define their intensity function as containing both an inhibiting and exciting component, as well as a constant background intensity and a term to deal with zero-inflation. In their simulation study (provided in their supplementary materials) they fix both of the latter to their true values, which only leaves two free parameters to be learnt by their ABC procedure. This means that they do not consider the full estimation problem where the background intensity needs to be learned in addition to the triggering kernel. To estimate the posterior distributions using ABC, Ertekin *et al.* (2015) use two summary statistics: the log-number of events and the Kullback–Leibler divergence between the histograms of the interevent times Δ_i of the true and simulated data set.

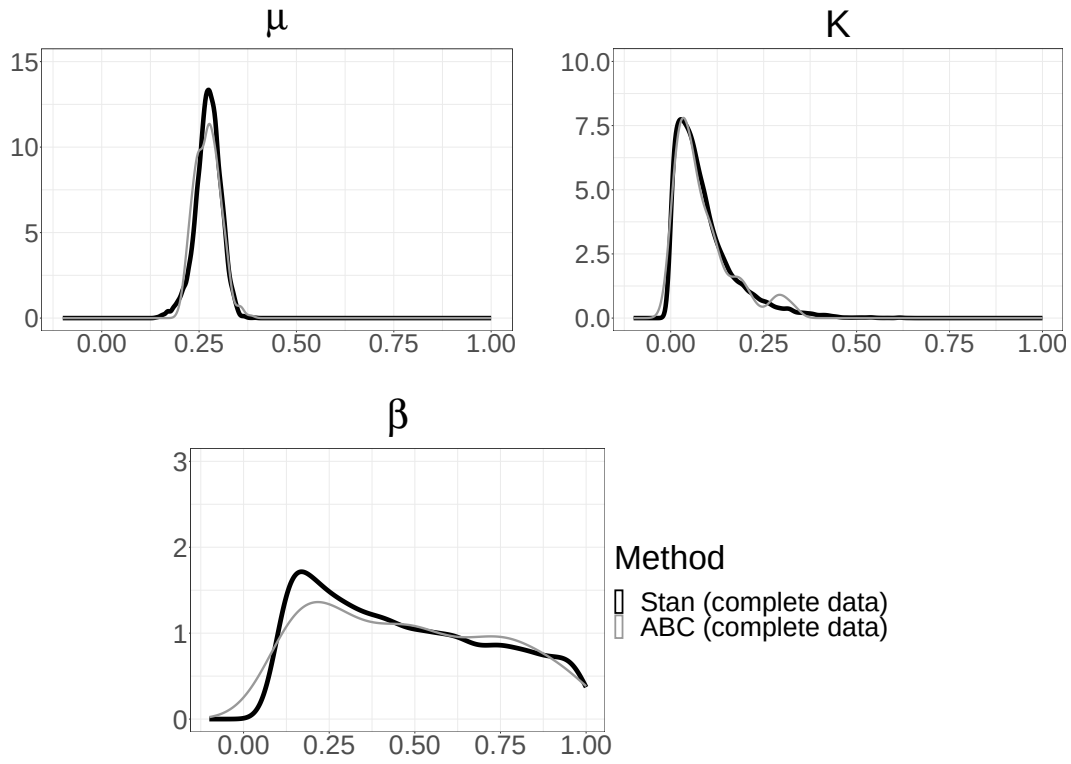


Figure 3.1: Undisturbed data posterior distributions for μ, K, and β. Parameters (μ, K, β) are chosen as $(0.2, 0.3, 0.3)$ and $T = 500$. The black curve shows the true posterior estimated by samples generated from Stan, grey represents ABC-Hawkes. All estimates are based on the complete, undistorted data sets.

For our comparison, we use a constant background intensity μ, which is assumed fixed and known for all methods to facilitate a direct comparison with Ertekin *et al.* (2015). Hence, we are left to estimate the posteriors of the two free parameters from the exponential excitation kernel, hence $\boldsymbol{\theta}_{\text{Ertekin}} = (K, \beta)$. The priors for the remaining Hawkes parameters (K, β) are chosen as in Section 3.4.1. Without any data distortion, Figure 3.2 and Appendix A.1.2 compare the true posterior distribution to the estimates using the two summary statistics from Ertekin *et al.* (2015) and ABC-Hawkes. It is evident that ABC-Hawkes does a better job at capturing the posterior distributions, showing that it can estimate the Hawkes parameters. This does not come as a surprise given the fact that Ertekin *et al.* (2015) only use two summary statistics when our proposed method utilises seven. In addition, Ertekin *et al.* (2015) make the assumption that the background rate μ is known, which is not the case for many applications. Our approach does not require this.

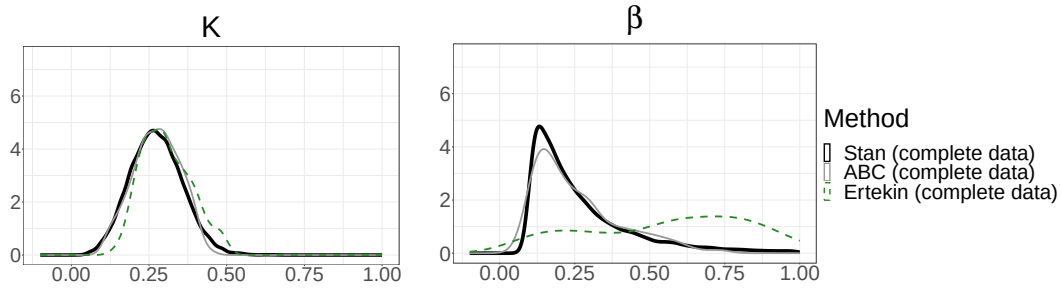


Figure 3.2: Undisturbed data posterior distributions for K and β. Parameters (K, β) are chosen as $(0.3, 0.3)$, $\mu = 0.3$ is fixed, $T = 500$. The solid black curve shows the true posterior estimated by samples generated from Stan, grey represents ABC-Hawkes, dashed green uses the summary statistics from Ertekin *et al.* (2015). All estimates are based on the complete, undistorted data sets.

3.4.2 Gap Deletion: Twitter/X Example

We now look at the first distortion scenario and apply the ABC-Hawkes algorithm to parameter estimation when there is a gap in the data (see Section 3.3.1) and compare our proposed approach to that of Tucker *et al.* (2019). For this purpose, we will manually insert distortion into an undistorted data set. This allows us to compute the posterior distributions of the original undistorted data, which serves as a reference. In this section, we examine the case when the start and end of the gap are known on a real Twitter data set. While Twitter has rebranded as "X" we use both names throughout this chapter.

We compare our approach to that proposed by Tucker *et al.* (2019). It is a tailor-made method to deal with a gap in the data, as showcased in their example on terrorism data. In that application, observations are missing since the physical index cards used to store the event were lost. Their algorithm is an MCMC-based method that treats the missing events as latent variables that are imputed at each iteration of the algorithm. It alternates between two steps. First, they sample

the missing data based on the complete pre-gap data and a given parameter vector. Second, they use the imputed events to construct a pseudo-complete data set, where the likelihood can be evaluated. Using this completed data set, they propose another parameter vector in classic Metropolis-Hastings fashion. This approach specialises in gaps in the data and cannot be readily employed when data is missing throughout the observational period.

For this example, we choose to study the occurrence time of tweets on Twitter/X, which previous research has shown can be accurately modelled by a Hawkes process (Mei and Eisner, 2017). We use a data set that was previously analysed by Rizoiu *et al.* (2017) and describes the retweet cascade of an article published in the New York Times. Note that for this example we do not need to consider events that happened before time 0, as we know none of them exist since no tweets can happen before the original news story was published.

While this data set is complete (and we can hence obtain the complete-data posterior distribution), we will manually create a gap in the data to assess whether the complete-data posterior can be recovered using only this distorted data. To evaluate ABC-Hawkes, we use the first 150 event times in the tweet data. To artificially create a gap, we delete all observations from observation t_{60} to observation t_{90}, to produce the incomplete data as shown in Figure 3.3. The relatively uninformative priors are chosen as in Section 3.4.1.

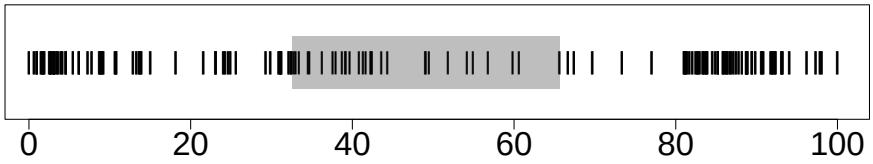


Figure 3.3: Twitter/X data observations. Each line indicates one event. Events inside the shaded area are removed to create the distorted data set.

We generate samples from the complete-data parameter posteriors using Markov Chain Monte Carlo as implemented in Stan (Stan Development Team, 2019) applied to the complete data. This represents complete-data posterior that would be obtained if we had access to the true undistorted data. We then apply ABC-Hawkes to the observed data only (i.e. the distorted, incomplete data), with the goal of recovering the complete-data posterior. In the implementation of the ABC-MCMC algorithm, we use independent random walk Gaussian proposal distribution for the $q(\cdot \mid \cdot)$ transition kernels, with standard deviations $(0.05, 0.05, 0.2)$ for (μ, K, β) respectively. To assess the performance of ABC-Hawkes, we compare the obtained posterior to two alternative methods that could be used: (1) MCMC (using Stan) applied to the incomplete data, which represents the naive attempt to learn the Hawkes parameters directly using only the observed data and ignoring the fact that data is missing. (2) The missing data algorithm suggested by Tucker *et al.* (2019) as described above.

We note that approach (2) is only applicable to this specific choice of distortion function where the distortion consists of no detected events at all during the gap period, and (unlike ABC-Hawkes), is not applicable to more general types of distortion, as will be discussed in the next section.

Alternatively, one could treat the two fully observed intervals as two realisations from the same process. However, when gaps are short and influence kernels are more heavy-tailed, events from the first period can have offsprings in the second period. Hence, we instead focus on methods that actively consider the missing data, such as ABC-Hawkes and the approach by Tucker *et al.* (2019).

Table 3.1: Twitter/X data posterior mean (and standard deviation)

Model	μ	K	β
Complete-Data Posterior	**0.55 (0.13)**	**0.65 (0.10)**	**0.91 (0.28)**
ABC-Hawkes	0.57 (0.16)	0.69 (0.11)	0.96 (0.47)
Naive	0.22 (0.08)	0.80 (0.07)	0.87 (0.23)
Tucker *et al.* (2019)	0.61 (0.14)	0.66 (0.10)	1.08 (0.34)

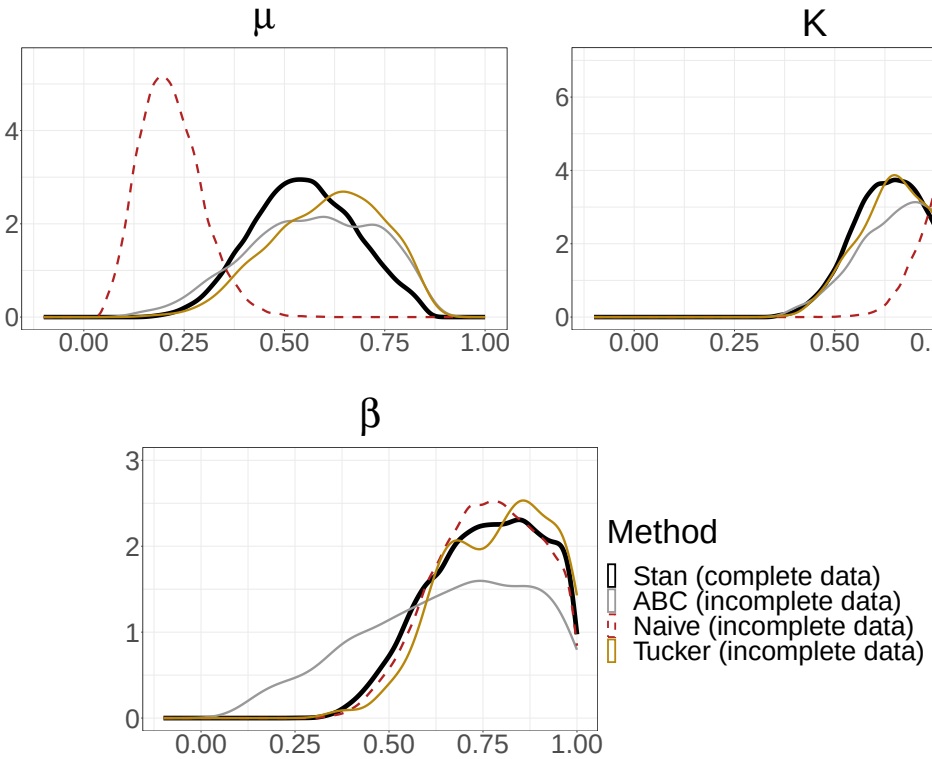


Figure 3.4: Twitter/X data posterior distributions for μ, K, and β. Solid represents the complete-data posterior using both observed and missing data estimated using Stan. Solid grey represents ABC-Hawkes using only the incomplete data. The naive approach using only the incomplete data is dashed red, while the solid yellow lines show the Tucker *et al.* (2019) imputation method.

Table 3.1 provides posterior means (with standard deviations) and Figure 3.4 shows the resulting posterior density estimates for all these methods. It can be seen that the ABC-Hawkes algorithm does a good job of recovering the posterior distribution despite the missing data, and is very close to the complete-data posterior mean for each of the model parameters. In contrast, the naive approach (which ignores the gap) produces a highly biased posterior which is not close to the complete-data posterior means of μ and K. The approach from Tucker *et al.* (2019) does substantially better than the naive approach with similar results to ABC-Hawkes. However, unlike ABC-Hawkes, Tucker *et al.* (2019) can only be employed when data is missing in a gap.

3.4.3 Constant Deletion

A key advantage of our approach is that it can also be used when the data is distorted in other ways. To investigate this, we use a simulation study where we generate data sets from a Hawkes process and manually distort them using constant deletion. The prior distribution for $\boldsymbol{\theta}$ is taken to be the same as above and we again used Stan (Stan Development Team, 2019) to sample from the idealised parameter posterior using the undistorted data, which acts as a ground truth that would be obtained if no distortion were present.

First, we investigate the bias introduced by distortion on a naive estimation that does not take distortion into account. We then examine both the case of known and unknown distortion parameters in the case of constant deletion, where each event has a probability of $\xi = a$ of being included in the data. Unlike in the above gap-missing data case, we are not aware of any other published algorithms which can handle this type of distortion, so the only other comparison we make is to the naive method which learns the posterior using the observed data without taking the distortion into account.

Simulation Study

The goal of this simulation study is to highlight the impact of missing data on a naive estimation that treats the incomplete data set as if it was complete. One could consider employing a naive maximum likelihood estimation and then rescaling the obtained parameter estimates according to the known a. However, we now showcase that the relation between the naive estimate and the true parameter is not simple and, in addition, heavily depends on the true parameter value.

To that extent, we simulate data sets where two parameters are fixed and one is varied. These data sets are then distorted with different levels of a ranging from 1 to 0.1 in increments of 0.1. Using these distorted data sets, we produce a maximum likelihood estimate for all three parameters for each data set. This is repeated 20 times for each such scenario. In Figure 3.5 we plot the mean estimate (and shaded 95% confidence interval) for the parameter that is not fixed.

This shows that deletion has a serious, complex effect on all parameter estimates when the distortion is not taken into account. While the effect is rather linear for μ, it is more intricate for K and β. Hence, rescaling naive estimates

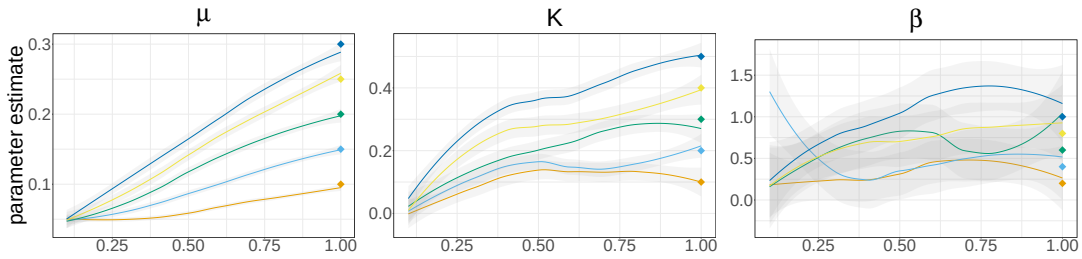


Figure 3.5: Mean maximum likelihood estimates for μ, K, and β as a function of a. We use five different sets of parameters per plot, each indicated by a coloured line. The parameter in the title of each plot is varied while the other two parameters are kept fixed. Coloured dots at $a = 1$ indicate the true parameter value per colour. The other ones are chosen as follows: **Left**: fixed $K = 0.3, \beta = 0.6$. **Middle**: fixed $\mu = 0.15, \beta = 0.6$. **Right**: fixed $\mu = 0.15, K = 0.3$. Shaded areas represent 95% confidence interval.

is not straightforward and would require additional research. This warrants a principled estimation approach that takes deletion into account.

ξ Known

First, we consider the distortion parameter a is assumed to be known. We simulate data with parameters $(\mu, K, \beta) = (0.2, 0.3, 0.3)$ and $T = 500$, and detect events with $a = 0.75$. In Figure 3.6 the posterior density estimates for $\boldsymbol{\theta}$. Additional examples can be found in Appendix A.1.3.

Again, ABC-Hawkes manages to learn the model parameters accurately and produces a posterior distribution which is close to the complete-data posterior, despite about a quarter of the data being deleted with $a = 0.75$. In contrast, the naive approach is severely biased, in particular for μ, and does not get close to the complete-data posterior. We note that the slight overestimation of the posterior variance is an inherent issue with ABC that comes from a necessary non-zero choice of ϵ_p (Li and Fearnhead, 2017).

ξ Unknown

As a final example, we investigate the performance of ABC-Hawkes under unknown, constant deletion. As above, we can only make a comparison to the naive approach, as, to the best of our knowledge, no other algorithm can explicitly handle this kind of distortion. Additional examples are available in Appendix A.1.4

For the distortion parameter $\xi = a$ we use the following vague prior:

$$a \sim \mathcal{U}(0.5, 1).$$

Figure 3.7 shows the posterior distributions with parameters $\boldsymbol{\theta} = (0.1, 0.5, 0.2)$ and $a = 0.65$. Even if the posterior of a is less accurate, ABC-Hawkes is reasonably successful in recovering the complete-data posterior for $\boldsymbol{\theta}$.

However, the estimation is difficult when the parameters are chosen in such a way that the observed data is close to that of a Hawkes process with a different

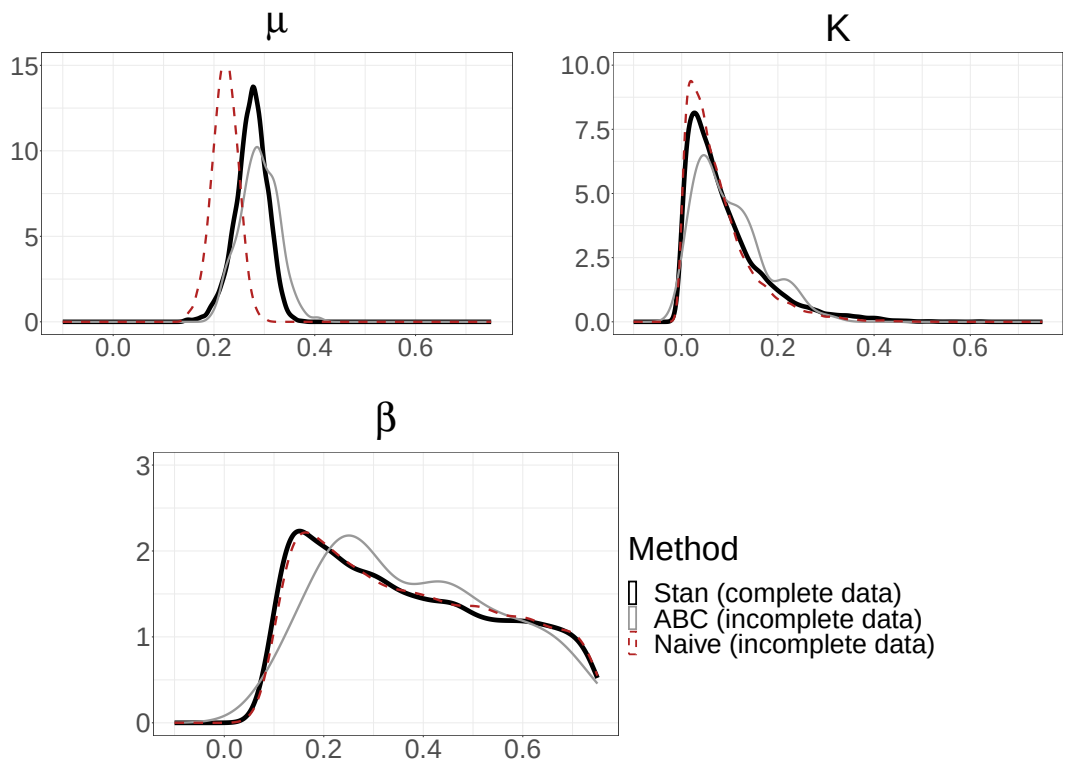


Figure 3.6: Data with known constant deletion posterior distributions for μ, K, and β . Parameters (μ, K, β) are chosen as $(0.2, 0.3, 0.3)$, $a = 0.75$, and $T = 500$. The solid black curve shows the complete-data posterior estimated by samples generated from Stan. Solid grey represents ABC-Hawkes, and the naive approach, both using only the incomplete data is dashed red.

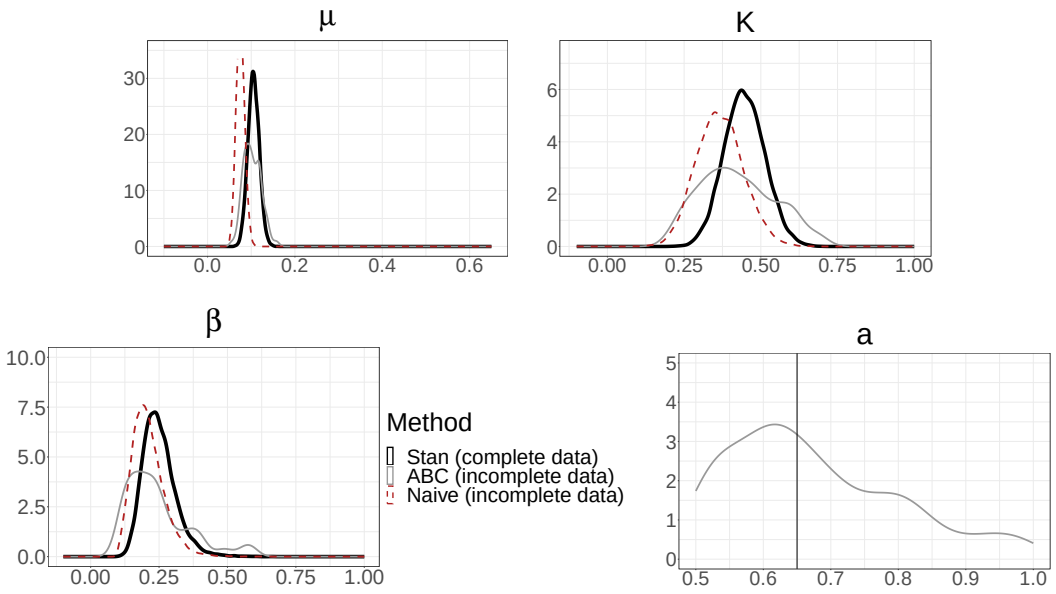


Figure 3.7: Data with unknown constant deletion posterior distributions for μ, K, β, and a. Parameters (μ, K, β) are chosen as $(0.1, 0.5, 0.2)$, $a = 0.65$, and $T = 2000$. The solid black curve shows the complete-data posterior estimated by samples generated from Stan. Solid grey represents ABC-Hawkes, and the naive approach, both using only the incomplete data is dashed red. The vertical line in the plot on the bottom right represents the true value.

parameter vector. We stress that this is not a shortcoming of our ABC-Hawkes procedure, but is an inherent issue relating to identifiablity, which requires considerations outside the scope of this work. Figure 3.8 showcases such a scenario. Here, without strong prior information, ABC-Hawkes cannot accurately recover the posterior distributions.

3.5 Discussion

In this chapter, we have demonstrated that it is possible to successfully learn the parameters of a Hawkes process even when the data is distorted. We have based our algorithm on ABC-MCMC, which has been adapted to the unique structure of a self-exciting point process. Unlike a naive MCMC approach which ignores the potential distortion, the resulting ABC-Hawkes algorithm can learn the complete-data posterior distribution that would have been obtained given access to the undistorted data. The strong performance of ABC-Hawkes was demonstrated using two realistic data-distorting mechanisms. Future research could expand the theory to other data-distorting mechanisms, for example, additive noise in a multivariate setting (Trouleau *et al.*, 2019), censoring (Xu *et al.*, 2017), binning (Shlomovich *et al.*, 2022), dependency on previous events such as undetected events after a large earthquake (Helmstetter *et al.*, 2006), or multiple gaps in the data. Venturing into ecology to address imperfect sampling and presence-only data (Renner *et al.*, 2015) would require an adaption to spatial processes.

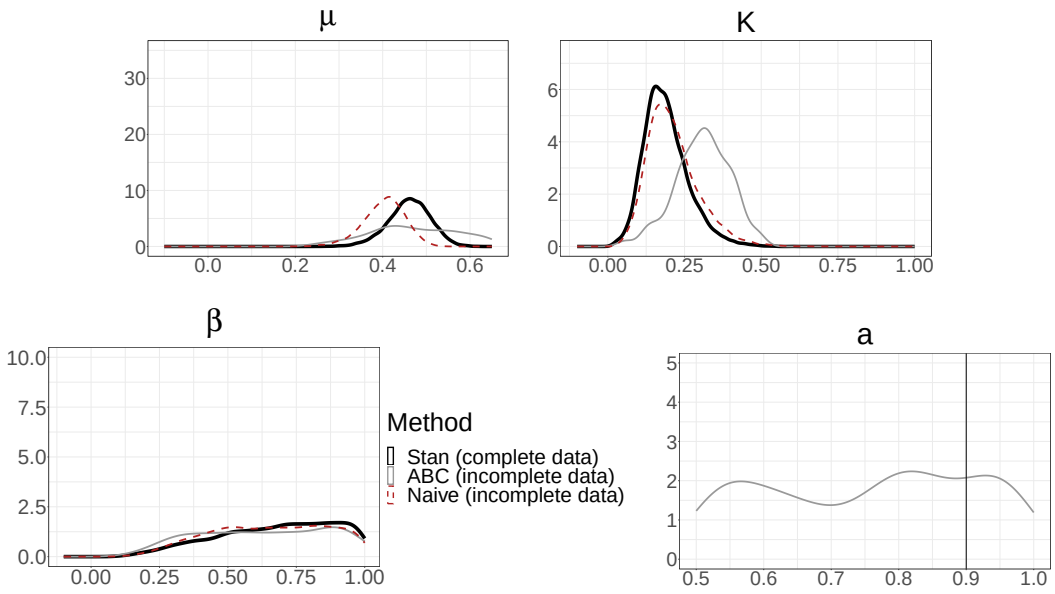


Figure 3.8: Data with unknown constant deletion posterior distributions for μ, K, β, and a. Parameters (μ, K, β) are chosen as $(0.4, 0.3, 0.8)$, $a = 0.9$, and $T = 500$. The solid black curve shows the complete-data posterior estimated by samples generated from Stan. Solid grey represents ABC-Hawkes, and the naive approach, both using only the incomplete data is dashed red. The vertical line in the plot on the bottom row represents the true value.

We could also consider the initialisation problem and related edge effects due to the unknown history of events before the observational period with $t < 0$ (Møller and Rasmussen, 2005), which can be seen as missing data. Only a few instances in the missing data literature mention this explicitly, e.g. Xu *et al.* (2017). While some of the proposed methods could not handle such a form of missing data (such as Tucker *et al.*, 2019), others might be able to accommodate this (e.g. Linderman *et al.*, 2017; Shelton *et al.*, 2018). We believe ABC-Hawkes could also be readily adapted to the initialisation problem.

We have also examined the constant deletion scenario highlighting difficulties that can arise. This limits not just ABC-Hawkes but any algorithm attempting to estimate posterior distributions in such scenarios. Other distortion scenarios, in particular those depending on data or covariates, are part of future work. Furthermore, it would be interesting to explore whether issues around identifiability can be resolved through specific choices of the influence kernel $g(\cdot)$ and distortion function $h(\cdot)$.

While our study focused on a simple Hawkes process with a parametrically specified self-excitation kernel, we seek to explore how our approach can be applied to other specifications of the Hawkes process as long as the resulting model is generative so that data sets can be simulated conditional on a parameter vector $\boldsymbol{\theta}$. This includes different background rates and excitation kernels, as well as specifications of the Hawkes process using nonparametric estimation (Chen and Hall, 2016) or LTSM networks (Mei and Eisner, 2017), which are potential avenues for future research.

Chapter 4

Hawkes Processes with Inhibition

So far we have assumed that K is non-negative. In this chapter, we now examine a multivariate Hawkes process when this requirement is relaxed and $K_{jm} < 0$ is allowed. Such a negative influence magnitude is called *inhibition*. This extension to the Hawkes process is subsequently used in Chapter 5 when constructing a model that can capture product cannibalisation. When $K_{jm} \leq 0$ then an event in dimension j decreases the intensity function of dimension m, hence making it less likely that an event in dimension m takes place.

Inhibition is a substantial extension to the general Hawkes process, which brings additional challenges and subtleties to the estimation procedure. When inhibition is present, the intensity function needs to be non-negative, as required by design. This has a knock-on effect on the computation of Λ, the integral of the intensity, which is needed for the likelihood evaluation. In addition, we also found that the commonly used conditions to assess stability were too restrictive under inhibition. We describe these issues in detail and propose an adapted condition that is less restrictive than the previously used ones. This chapter investigates intricacies regarding the non-negativity of the intensity, the integral of the intensity, and stability, while Chapter 5 puts these considerations into practice.

Hawkes process models with inhibition have been used in a variety of applications. A prominent application of Hawkes processes with inhibition can be found in neural spike trains, where the inhibition captures the period of decreased activity after a neural spike event (Eichler *et al.*, 2017). Moreover, such models have been successfully employed in finance (Lu and Abergel, 2018; Tetereva, 2018) or in adequately modelling the popular MemeTracker data set (Lemonnier and Vayatis, 2014).

This chapter is structured as follows. Section 4.1 reviews issues around the required non-negativity of the intensity function. In Section 4.2 different approaches to compute the integral of the intensity, which is required for the likelihood, are presented. Finally, Section 4.3 reviews current conditions for stability and proposes a new, less restrictive one when inhibition is present.

4.1 Non-Negative Intensity

One concern of Hawkes processes with inhibition is the necessity of a non-negative intensity function. Any point process requires, by design, that the intensity function at every $t \in [0, T]$ is non-negative. When a K_{jm} is negative in the previously seen intensity function

$$\lambda_m(t) = \mu_m + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i = j, \\ t_i < t}} K_{jm}\, g_{jm}(t - t_i),$$

it is not guaranteed that the intensity always stays non-negative. We discuss two approaches on how to handle this issue in Section 4.1, as well as the implication for the likelihood in Section 4.2. There are two main approaches in the literature to ensure non-negativity: restricting the parameter space and using a link function to ensure positivity.

4.1.1 Restricting the Parameter Space

One – somewhat crude – way to guarantee a non-negative intensity function is to restrict the parameter space based on the observed data set such that the intensity always stays non-negative. However, this method has three fundamental shortcomings: potential non-consistency, sampling issues, and data dependency.

1. It is typical that the intensity function of a Hawkes process exhibits many spikes and drops with short, rapid changes. Excluding all parameter combinations that lead to a decrease below zero *somewhere* would substantially restrict the parameter space, where many parameter combinations are prohibited. This could potentially exclude the true parameters if inhibition is truly present, which would lead to a non-consistent estimation procedure.

2. Restricting the parameter space can cause sampling issues as MCMC samplers are prone to boundary artefacts under such confining conditions.

3. The 'permissibility' of a parameter combination depends on the data. A set of parameters causing a non-negative intensity on a data set does not guarantee this property when more data is collected.

We therefore do not recommend this approach and instead rely on a link function, as described next.

4.1.2 Link Function

Since restricting the parameter space to ensure a non-negative intensity function has clear drawbacks we instead use a link function $\phi(\cdot)$. This is another common practice and leads to the intensity

$$\lambda_m(t) = \phi\left(\mu_m + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i = j, \\ t_i < t}} K_{jm}\, g_{jm}(t - t_i)\right). \tag{4.1}$$

The link function $\phi(\cdot) : \mathbb{R} \to \mathbb{R}^+$ ensures a non-negative intensity at every t. For example, Mei and Eisner (2017) use the *softplus* function $\phi(x) = s \log(1 + \exp(x/s))$ with parameter s. Another straightforward choice of ϕ is the $ReLU$ function where

$$\phi(x) = \max(a, x), \tag{4.2}$$

for a small, non-negative a. A popular approach in the literature is to set $a = 0$ (Lemonnier and Vayatis, 2014; Lu and Abergel, 2018; Costa *et al.*, 2020), as this also preserves the interpretation of K_{jm} as the average number of offsprings when K_{jm} is nonnegative. In Chapter 5 we use the $ReLU$ function with $\phi(x) = \max(0, x)$.

4.2 Integrating the Intensity

Many parameter estimation methods, such as our subsequent Bayesian procedures, need to evaluate the likelihood, which requires $\Lambda = \int_0^T \lambda(t)\, dt$. We now discuss different approaches from the literature, present exact solutions for particular choices of the influence kernel, and suggest an approximate solution that will be subsequently utilised.

When all K_{jm} are non-negative and hence no link function is needed, the integral of the intensity can be computed by integrating each segment between events (as well as the ones between 0 and the first event and between the last event and T). This reduces to an easy computation of the integral that sums over the background rate and the contributions of each event (Ogata, 1981), as seen in Section 2.3. However, it becomes difficult to integrate the intensity function under inhibition in this scenario as the intensity might drop below zero. Lu and Abergel (2018) choose to calculate the integral as in the excitation-only case, even though the parts of the integral below zero contribute negatively to the integral.

The use of a link function $\phi(\cdot)$ guarantees a non-negative intensity function under inhibition (see Section 4.1.2). However, any choice of link function that is not the identity prohibits a straightforward calculation of the integral as the approach by Ogata (1981) cannot be employed anymore.

Mei and Eisner (2017) explain in their supplementary materials that they approximate Λ for a one-dimensional Hawkes process by sampling a single t^* uniformly from $[0, T]$ and then use $\hat{\Lambda} = T\lambda(t^*)$. This $\hat{\Lambda}$ approximates the average intensity and is indeed an unbiased estimator for Λ, but has a large variance. The intensity functions often exhibit rapid ups and downs, such that just evaluating it once cannot capture all aspects of the function. Ertekin *et al.* (2015) use Approximate Bayesian Computation to circumvent the problem as this method does not require an evaluation of the likelihood and therefore the integral does not need to be calculated.

In the subsequent sections, we present two alternative approaches. Integrating the intensity of a Hawkes process with $\phi(x) = \max(0, x)$ is equivalent to identifying the intervals of the intensity function without a link function that are non-negative and integrating only those (Bonnet *et al.*, 2021). While this is an exact solution, this approach requires finding the roots of the intensity function (without a link function) for an exponential kernel. We show how one can find

the exact value of Λ by identifying the roots. This is a computationally expensive endeavour. In addition, we propose an approximation procedure of Λ based on Simpson's rule for any parametric kernel. The latter is subsequently employed in Chapter 5.

4.2.1 Exact Solution

When using a ReLU link function $\phi(x) = \max(0, x)$, the integral of the intensity with link function is equivalent to the sum of non-negative parts of the intensity without a link function (Bonnet *et al.*, 2021). Hence, we now investigate root-finding approaches for the intensity function (without link function). We first show that the roots of an intensity function (without a link function) with the exponential kernel are the solutions to a high-order polynomial. The computational complexity of this root-finding approach is high since the root-finding needs to be employed between each event time in every dimension. When setting all β_{jm} to be equal, we find a simple expression for the roots when using the exponential kernel. However, this assumption may be too restrictive in application.

General Case

Bonnet *et al.* (2021) provide an exact integral for the one-dimensional case. We look at the multivariate Hawkes process and find an exact solution for its roots without placing any restrictions on the parameters. It turns out that this is a polynomial problem.

Assume that we have data Y from M dimensions and the intensity function is defined with the exponential kernel. The intensity can only drop below zero when an event happens. Therefore it is sufficient to check only intervals after an event at which the intensity is negative to see if the intensity becomes positive again before the next event happens. When $\lambda_m(t_n + \varepsilon) = 0$ for $\varepsilon \to 0$ then we can find the root t' between observation t_n and t_{n+1} in the following way:

$$\mu_m + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i=j, \\ t_i<t'}} K_{jm}\, \beta_{jm}\, \exp\left(-\beta_{jm}(t' - t_i)\right) = 0 \tag{4.3}$$

$$\mu_m + \sum_{j=1}^{M} x^{-\beta_{jm}} \underbrace{\sum_{\substack{i:\, d_i=j, \\ t_i<t'}} K_{jm}\, \beta_{jm}\, \exp\left(-\beta_{jm} t_i\right)}_{v_{jm}} = 0 \text{ where } x = \exp(t') \tag{4.4}$$

$$\mu_m + \sum_{j=1}^{M} x^{-\beta_{jm}} v_{jm} = 0. \tag{4.5}$$

This is now a polynomial in x that needs to be solved, which can be computationally expensive, in particular since this needs to be conducted in each dimension each time the intensity has dropped below zero due to an event. However, this naive method could be improved through the recursive representation of the Hawkes likelihood (Ozaki, 1979; Laub *et al.*, 2021).

Restrictive Case

Here we examine a special case where all β_{ij} are equal and find the roots for a M dimensional Hawkes process. Note that Bonnet *et al.* (2023) give similar results (though they use $\beta_{ij} = \beta_j$), which was first submitted to ArXiv in May 2022. However, an earlier version of this work (including the results below) was available on the ArXiv from January 2022 onward.

We use the same setup as above where we assume data Y from a M dimensional Hawkes process. Under the restriction $\beta_{ij} = \beta$ it is easier to find the root t' between observation t_n and t_{n+1} when $\lambda_m(t_n + \varepsilon) = 0$ for $\varepsilon \to 0$:

$$\mu_m + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i=j, \\ t_i<t'}} K_{jm}\, \beta\, \exp\left(-\beta(t' - t_i)\right) = 0 \tag{4.6}$$

$$\exp\left(-\beta t'\right) \underbrace{\left[\sum_{j=1}^{M} \sum_{\substack{i:\, d_i=j, \\ t_i<t'}} K_{jm}\, \beta\, \exp\left(\beta t_i\right) \right]}_{v} = -\,\mu_m \tag{4.7}$$

$$t' = \frac{\log\left(\frac{-\mu_m}{v}\right)}{-\beta}. \tag{4.8}$$

if $t_n < t' < t_{n+1}$. However, this condition on β_{ij} might be too restrictive for real-life applications.

4.2.2 Approximation

Instead, we propose using a numerical approximation of Λ without placing any restrictions on β_{ij}. In each segment between events in every dimension, we use a cubic Simpson's rule approximation. While this is a costly approximation in terms of computational complexity, we found that any lower-level approximation introduced too much bias. We use the Cubic Simpson's Rule, such that

$$\int_a^b f(x)\, dx \approx \frac{(b-a)}{8} \left[f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b) \right]. \tag{4.9}$$

Assume M-dimensional data Y and intensity function $\lambda_m(\cdot)$ for each dimension $m = 1 \dots M$. Algorithm 2 outlines the resulting approximation procedure used.

4.3 Stability

Due to the self-exciting behaviour of the Hawkes process, it is possible that an infinite number of events take place in finite time. For example, this can happen in a one-dimensional Hawkes process if each event has more on average one or more offsprings. This behaviour is called *supercritical* (Helmstetter and Sornette, 2002) or *explosive* (Browning *et al.*, 2021). However, for real-life applications, it

Algorithm 2 Approximating Λ using the Simpson's Rule

Define $\mathcal{H}_t = \{t_i : t_i \leq t\}$
Set res $= 0$
for i in $0 : N$ **do**
  if i $=$ 0 **then**
    Set $a = 0$
  else
    Set $a = t_i$
  end if
  if i $=$ N **then**
    Set $b = T$
  else
    Set $b = t_{i+1}$
  end if
  for m in $1 : M$ **do**
    Set res $=$ res$+\frac{(b-a)}{8}\left[\lambda_m(a \mid \mathcal{H}_a) + 3\lambda_m\left(\frac{2a+b}{3} \mid \mathcal{H}_a\right) + 3\lambda_m\left(\frac{a+2b}{3} \mid \mathcal{H}_a\right) + \lambda_m(b \mid \mathcal{H}_a)\right]$
  end for
end for
Return res as the approximation of Λ

can be desirable to limit the parameter space to non-explosive instances (Kolev and Ross, 2019). This is referred to as *stability* (Bremaud and Massoulie, 1996; Bacry *et al.*, 2020).

Definition 4.3.1. A Hawkes process with intensity $\lambda(t)$ is stable if there exists a unique stationary distribution of the process where

$$\sup_{t\in[0,\infty)} \lambda(t) < \infty, \tag{4.10}$$

i.e. the intensity function is finite. (Hawkes and Oakes, 1974; Bremaud and Massoulie, 1996; Bacry *et al.*, 2015).

When using Hawkes processes one usually assumes, either explicitly or implicitly, that the process is stable. This ensures that there are not infinitely many events happening. In Bayesian statistics, this can be easily accomplished by restricting the prior space. However, we found that the two stability conditions currently used in the literature are unnecessarily restrictive when inhibition is present. We therefore propose a new condition which is less restrictive than the two existing ones, which can be used when specifying a prior. This section contains the derivation of our condition and compares it to the two existing ones.

For a univariate Hawkes process, stability is achieved by restricting $K < 1$. This relates to the branching interpretation described in Chapter 2.2.1, as this restriction ensures that each event has, on average, less than one offsprings. Hence, any cascade will "die out" eventually. In the excitation-only multivariate case, a Hawkes process is stable if the spectral radius, i.e. the largest absolute eigenvalue, of $\mathbf{K}$ is less than 1.

Two conditions (**C1**, **C2** defined below) have been used in the literature to determine stability in the multivariate case with inhibition. In Section 4.3.2 we introduce a new condition that is less restrictive than both of the previously used ones. We prove that when at least one of (**C1**, **C2**) holds, so does our condition. Moreover, there exist parameters $\mathbf{K}$ for which our condition holds when neither of (**C1**, **C2**) do. This permits us to have a unified approach for checking stability and to classify a larger set of parameters as stable. We also provide a toy example to illustrate the usefulness of our suggested condition.

We introduce the following notations: $\text{abs}(\mathbf{A})$ is an $M \times M$ matrix where each entry is $\mid A_{ij} \mid$, the absolute value of A_{ij}. Moreover, $\mathbf{A}^+$ is the matrix with entries $\max(A_{ij}, 0)$. We write $\rho(\mathbf{A})$ for the spectral radius of matrix $\mathbf{A}$, i.e. the largest absolute eigenvalue of matrix $\mathbf{A}$.

4.3.1 Stability Conditions in the Literature

To start, we state two conditions to assess stability for the multivariate Hawkes that have been introduced in the literature:

C1 A Hawkes process is stable if the spectral radius $\rho(\text{abs}(\mathbf{K})) < 1$ (Bremaud and Massoulie, 1996).

C2 A Hawkes process is stable if $\max_j \sum_{i=1}^{M} K_{ij}^+ < 1$ (Sulem *et al.*, 2021).

Note that **C1** uses the spectral radius of the absolute value matrix, whereas **C2** utilises only the positive part of the matrix $\mathbf{K}$. Both conditions are sufficient, but not necessary. Sulem *et al.* (2021) discuss both **C1** and **C2**, but do not compare them as neither is stronger than the other. As **C1** uses the absolute value matrix, negative entries (inhibition) are converted into excitation. Hence, while the intensity is *decreased* by certain events, stability is checked as if those events *increased* the intensity. Of course, this procedure is sufficient, but the intensity $\lambda_m(\cdot \mid \text{abs}(\mathbf{K}))$ may be rather different than the original $\lambda_m(\cdot \mid \mathbf{K})$, in particular when strong inhibition is present. Hence, we found there to be scope to develop a new criterion that is more closely tailored to processes that contain inhibition.

4.3.2 Introducing a New Condition

We propose a new condition to assess stability for a given parameter $\mathbf{K}$.

Theorem 1 (**C3**). *If the spectral radius $\rho(\mathbf{K}^+) < 1$, then the process with intensities $\lambda_m(\cdot \mid \mathbf{K})$, $m = 1 \dots M$, is stable.*

To prove it, we state an auxiliary lemma.

Lemma 2. *Suppose that for a process with intensity $\lambda^*(\cdot)$ it holds that*

$$\sup_{t \in [0,\infty)} \lambda^*(t) < \infty,$$

i.e. it is finite (Hawkes and Oakes, 1974). Now consider another process with intensity $\lambda(\cdot)$ such that $\lambda(t) \leq \lambda^*(t)$ for all $t \geq 0$. Then

$$\sup_{t\in[0,\infty)} \lambda(t) < \infty.$$

holds as well.

Proof. Since $\lambda(t) \leq \lambda^*(t)$ for all $t \geq 0$ this means that

$$\sup_{t\in[0,\infty)} \lambda(t) \leq \sup_{t\in[0,\infty)} \lambda^*(t) < \infty,$$

and hence

$$\sup_{t\in[0,\infty)} \lambda(t) < \infty$$

holds as well, i.e. $\lambda(\cdot)$ is finite. □

Using this lemma we now prove Theorem 1.

Proof. Since $K^+_{ij} \geq K_{ij}$ for each $i, j = 1 \dots M$, it follows that $\lambda_m(t \mid \mathbf{K}^+) \geq \lambda_m(t \mid \mathbf{K})$ for all $m = 1 \dots M$ and all $t \in [0, T]$. By Lemma 2, if a process with intensity $\lambda_m(\cdot \mid \mathbf{K}^+)$ is finite then a process with intensity $\lambda_m(\cdot \mid \mathbf{K})$ is finite as well. A multivariate process is finite if the intensity in each dimension $m = 1 \dots M$ is finite. If $\rho(\mathbf{K}^+) < 1$ then the multivariate process with intensities $\lambda_m(\cdot \mid \mathbf{K}^+)$, $m = 1 \dots M$, is finite, and therefore the multivariate process with intensities $\lambda_m(\cdot \mid \mathbf{K})$, $m = 1 \dots M$, is finite as as well. Uniqueness is guaranteed by Theorem 6.55 from Liniger (2009), and hence the process with intensities $\lambda_m(\cdot \mid \mathbf{K})$, $m = 1 \dots M$ is stable. □

4.3.3 Comparison

We now compare **C3** to both existing conditions (**C1**, **C2**) and show that if at least one of them holds, so does **C3**. Moreover, there are examples where only **C3** holds. This implies that **C3** can confirm stability for more parameters, which is useful when fitting a multivariate Hawkes process.

Theorem 3. *When **C1** holds, then **C3** holds as well.*

Proof. First, we compare **C3** to **C1** when all entries K_{ij} are non-negative (i.e. excitation only). It is trivial to see that when **C1** holds, then **C3** holds as well since $\text{abs}(\mathbf{K}) = \mathbf{K} = \mathbf{K}^+$ and therefore $\rho(\mathbf{K}^+) < 1$.

When we do not restrict the entries K_{ij} be non-negative, we note that each entry of $\text{abs}(\mathbf{K})$ is as least as large as the corresponding entry in $\mathbf{K}^+$. Within the entry-wise positive matrices, the spectral radius is monotonous (Lemma 12, p. 153 Serre, 2002). Hence, if $\rho(\text{abs}(\mathbf{K})) < 1$ then also $\rho(\mathbf{K}^+) < 1$.

Therefore, when **C1** holds, then **C3** holds as well. □

Theorem 4. *When **C2** holds, then **C3** holds as well.*

We first state the following lemma.

Lemma 5. *Let X be a $N \times N$ matrix with entry X_{ij} in row i and column j. Then $\rho(X) \leq \max_j \sum_{i=1}^{N} X_{ij}$.*

This is a direct consequence of the Gelfand formula (Gelfand, 1941).

With that, we can now provide the proof for the stated Theorem 4.

Proof. By Lemma 5, if $\max_j \sum_{i=1}^{N} K_{ij}^{+} < 1$ then also $\rho(\mathbf{K}^{+}) < 1$ and therefore **C3** holds if **C2** holds. □

Hence, we have shown that if at least one of (**C1**, **C2**) holds, so does **C3**. In addition, there are examples where neither of the existing conditions could confirm stability, but by using **C3** we can verify that the process is stable. Let us examine a two-dimensional Hawkes process with

$$\mathbf{K} = \begin{pmatrix} 0.5 & 1 \\ -2 & 0.5 \end{pmatrix}$$

as an illustrative example. Note that $\rho(\text{abs}(\mathbf{K})) > 1$ and the maximum column sum of $\mathbf{K}^{+}$ is also larger than 1, hence neither **C1** nor **C2** hold. Hence, by just using the two existing conditions it is not possible to assess whether a process using $\mathbf{K}$ would be stable. However, we can make use of **C3**, as $\rho(\mathbf{K}^{+}) < 1$ and confirm that a process using $\mathbf{K}$ is stable.

Note that **C3** essentially considers stability for a process without any inhibition. Hence, this criterion cannot determine stability for a process that is stable *because* of its inhibition, i.e. stable with inhibition but not stable without inhibition. Investigating this class of Hawkes processes and their properties constitutes an interesting avenue for future research.

In summary, **C3** not only provides a unified approach to assess stability, but it also permits us to determine stability for more parameters than by just using (**C1**, **C2**).

Chapter 5

Hawkes Processes for Product Cannibalisation

This chapter delves into the application of Hawkes processes to product cannibalisation. Based on our exposition in Chapter 4 we employ a Hawkes process with inhibition to capture product cannibalisation. We propose a new parametrisation for the influence magnitude as this permits a prior choice irrespective of the dimensionality M. Our suggested model is then fitted on real-life data and outperforms other models without inhibition.

Section 5.1 gives an overview of product cannibalisation, including a review of the literature. In Section 5.2 the data from our industry partner is described. Section 5.3 motivates a Hawkes process model for product cannibalisation. The estimation procedure is outlined in Section 5.4. Section 5.5 discusses the prior choice, which includes a prior based on a reparameterisation which is agnostic to the dimensionality of the problem. Two examples on real data are presented in Section 5.6. We finish with a discussion of our work in Section 5.7.

5.1 Product Cannibalisation

The purchase of one product can influence the sales volume of others in the same product category. For example, a consumer may choose to buy one particular good instead of a different (yet similarly designed) one. Such an effect is known as *product cannibalisation*. More formally, product cannibalisation in the marketplace is defined as the decrease in sales of one product due to the sales of a closely related product, or to the introduction of a new, similar product (Copulsky, 1976). We propose using a multivariate Hawkes process with inhibition to estimate product cannibalisation in a Bayesian framework.

Multiple approaches have been suggested to describe the dependence structure between product sales. A popular model comes from Ruiz *et al.* (2020) who consider the sequential choices of individual shoppers and how items can interact with each other. Other authors focus on similar articles that differ mainly in their quality (Desai, 2001). In this context Ghose *et al.* (2006) examine the impact of used books on online book sales. At each book purchase, a consumer can opt for a brand-new item or a pre-owned one. For example, it is of interest at what price

point a consumer switches to buying a used version of a title, hence cannibalising the sales of new books. Okorie *et al.* (2021) provide a meta-analysis for product cannibalisation focusing on re-manufactured goods in the circular economy.

Several of the examples mentioned above use simple models driven by marketing theory (Atasu *et al.*, 2010) or game theory (De Giovanni and Ramani, 2018). A different approach comes from Kamakura and Srivastava (1984), who include product cannibalisation in their probabilistic choice models that try to estimate the utility of each article. More recently, Guidolin and Guseo (2020) employ the Lotka-Volterra equations, a pair of non-linear differential equations commonly used in predator-prey scenarios, to capture the sales of the Apple iPhone, once the iPad was introduced. This models potential cannibalisation, e.g. a person buys an iPad rather than an iPhone or vice versa. Notably, this approach allows for asymmetric competition, e.g. more iPhone sold do not lead to less sales of iPads, but more iPad sales do lead to fewer iPhone sales. Kong (2015) use a logit regression to model the sales of an article depending on a variety of covariates, such as availability and display inside the store. They also include product cannibalisation (based on price and similarity) in their model.

The literature focuses on product cannibalisation for cumulative sales numbers instead of taking the temporal nature of (repeated) purchases into account. In recent years, a few notable exceptions have appeared. For example, Aguilar-Palacios *et al.* (2021) use a causal time series model to quantify product cannibalisation in grocery sales. Machine learning techniques have also recently been employed in this field of study. Bekal and Bari (2021) use boosting to predict sales in the presence of product cannibalisation, while Garnier (2022) relies on neural networks to uncover the "competition" between time series.

Most approaches use *retail* data, i.e. records of when a member of the public buys a product (US Census Bureau, 2011). For example, Garnier (2022) records when an individual consumer purchased a particular freezer. However, through our industry partner, we only have access to data on a *wholesale* level. Wholesale trade is defined as selling goods, often in large quantities, to other businesses who then might sell to the consumer for retail distribution (US Census Bureau, 2011). We refer to them as *wholesale customers*.

In contrast to the above approaches, we propose to model cannibalisation using a multivariate Hawkes process, where each product is represented by one dimension, with the sales being the events. Hawkes processes have been successfully used to flexibly model purchasing behaviour (see, for example, Pitkin *et al.*, 2018), but to the best of our knowledge, it is the first time that they are employed to estimate product cannibalisation.

To estimate product cannibalisation we use a multivariate Hawkes process where each article is represented by a dimension. An event describes a purchase by a wholesale customer where the respective article was included. This allows us to quantify the influence of one event (order placed for article i) on the probability of an event occurring in all dimensions (orders for article $j = 1 \dots i \dots M$). In particular, we investigate the case where an event does not make it *more* likely for another one to happen (excitation) but makes it *less* likely to occur (inhibition). This framework then allows us to capture product cannibalisation in the following manner. When an event in dimension i makes it less likely that an event happens

in dimension j, the sale of article i makes sales of article j less likely, which we interpret as product cannibalisation.

5.2 Data

We are collaborating with a major international company, anonymised as *CompanyCo*, who have given us access to some of their data. We are not able to publish the company's name or its industry due to non-disclosure agreements. This data is recorded on a wholesale level. Hence, it is not possible to infer when an individual consumer bought an item, but we know when a wholesale customer has placed an order with CompanyCo containing a certain article. For ease of exposition, we have summarised our terminology in Table 5.1, which includes an example from an unrelated industry.

Description	**Anonymised as**	**Unrelated Example**
Company	CompanyCo	The Coca Cola Company
Article = Product	Article 1, Article 2...	Coke Zero 8x330ml, Sprite 8x330ml, Costa Coffee Latte 250ml
Label	SomeLabel	Coke Zero, Sprite, Costa
Product Category	Product Class A Product Class B	Soft Drinks, Coffee
Wholesale Customer	BusinessGroup	Tesco, Sainsbury's
Individual Consumer	—	a person in a Tesco store

Table 5.1: Clarification of terminology, which includes an unrelated example from the beverage industry.

Each order placed by a wholesale customer at CompanyCo is recorded as one line in the data and consists of a certain quantity of exactly one article. If multiple articles are ordered at the same time, then there will be distinct entries per article. Each entry contains information on the article, the wholesale customer, and the order itself (e.g. the date on which the order was placed, and the quantity ordered). We do not have access to the price paid by the wholesale customer, but the data does contain the suggested retail price for each article. While this is not necessarily the price an individual consumer pays in the store (for example, due to special offers), it at least indicates which products may be more expensive/cheaper than others.

For this analysis, we focus on the arrival of orders without additional covariates. These order dates are recorded in discrete time where only the day of the order is known. However, it was desirable for us to make use of research and programming infrastructure from continuous-time point processes. Hence, we took the decision to translate our event times from discrete to continuous time stamps. To that end, we add random noise with distribution $\mathcal{U}(0, 1)$ to each event. This avoids estimation problems with interevent times that are always a multiple of 1, which would have appeared if the discrete event times were used without adjustment in a continuous model. To avoid introducing spurious excitation between

orders that happened simultaneously in the discrete data set, we add the same noise to those events. For example, if the event times were recorded as $(1, 4, 4)$, then the adjusted data set may look like $(1.5935, 4.1104, 4.1104)$.

The products sold by CompanyCo are goods used by individual consumers. Items that differ in appearance, but otherwise have the same design, are classified as different articles. There are also a multitude of labels within the CompanyCo brand that differ in price point or cater to particular consumer groups. One example of such a label is anonymised as *SomeLabel*.

For this analysis, we examine two categories of products from CompanyCo's portfolio, which we call Product Class A and Product Class B. For our application we will focus on the orders placed by one wholesale customer anonymised as *BusinessGroup*. This particular wholesale customer was chosen by CompanyCo for in-depth analysis due to their medium-sized and limited purchasing power such that they need to make active selections on which articles to stock.

First, we examine when BusinessGroup is placing orders that contain products from one product class with CompanyCo. This is the same product class examined in Section 5.6.2. In Figure 5.1 we plot the orders placed for eight articles over the course of 18 months. We observe a distinct seasonality in the orders of BusinessGroup. According to CompanyCo this is a typical pattern driven by the preorders and reorders of goods in a year, which is prevalent in most wholesale customers.

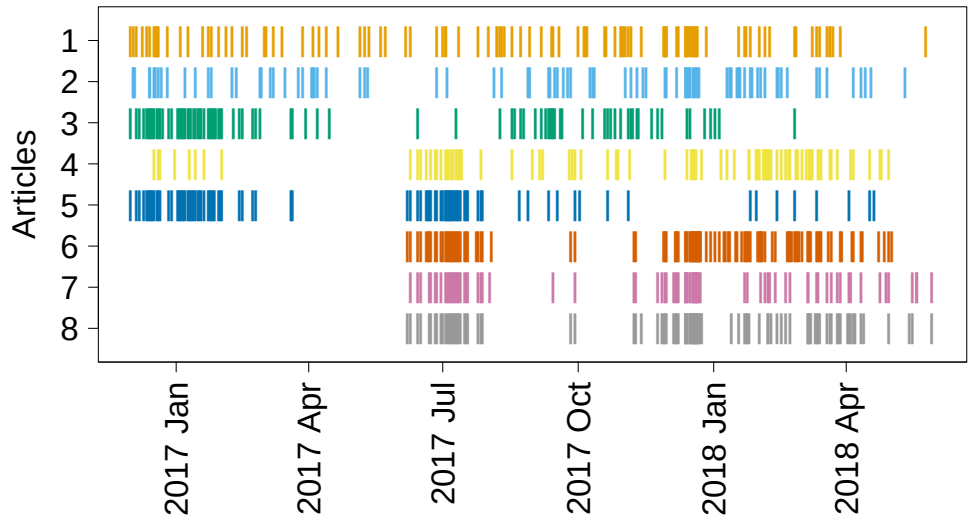


Figure 5.1: Orders placed for eight products from Product Class B by BusinessGroup. This represents a subset of all products in this class. Each vertical bar indicates that on the particular day an order was placed that included the respective article.

Three patterns can be observed: monthly and weekday seasonality, and a separate Christmas period (defined as 24^{th} to 27^{th} of December). We now examine each of these visually. Figure 5.2 displays the number of orders for products in Product Class B placed by BusinessGroup per month over 1.5, years which gives

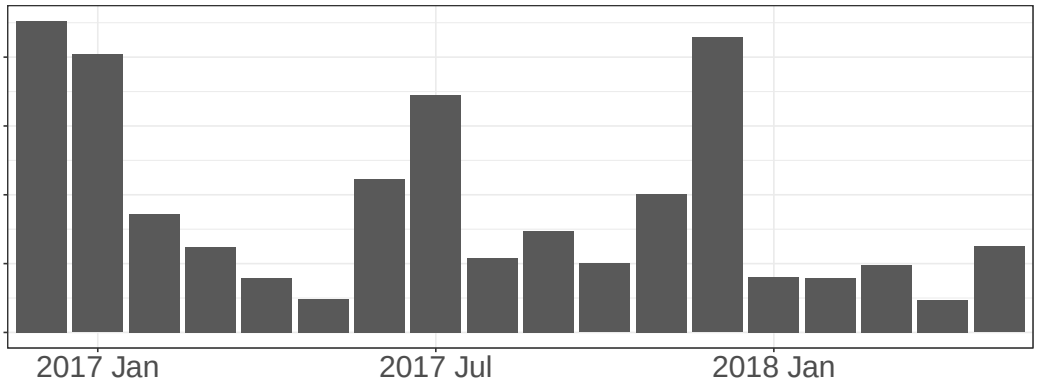


Figure 5.2: Orders placed each month for all products in Product Class B by BusinessGroup. The y-axis has been removed for data protection reasons.

a clear indication for a monthly variation. Figure 5.3 plots the daily number of orders for December 2016 and January 2017. The plot shows that the number of orders also varies greatly between weekdays. In addition, there is a sharp drop in orders over the Christmas period, irrespective of the weekday and the otherwise large order numbers in December, likely due to bank holidays. Section 5.4 describes how these three characteristics are utilised in the background rate of our model. This seasonal pattern is also prevalent across product classes.

We now examine individual articles to motivate our search for product cannibalisation. Figure 5.4 plots the arrival of orders by BusinessGroup for two similar products. Apart from the seasonal variability two trends are visible. Firstly, orders for each article are placed in rapid succession of themselves, i.e. when Article 1 is ordered it becomes more likely that the same article will be ordered again soon. The same holds true for Article 2. This *self-excitation* is layered on top of the seasonal trends discussed above. Secondly, the point process displays *cross-inhibition* where the order of Article 1 makes it less likely that Article 2 is purchased. Interestingly, this relationship also holds in reverse where a purchase of Article 2 makes it less likely that Article 1 is ordered. Given the similarities of the two articles (appearance, suggested retail price, no particular label) it is sensible to interpret this inhibition as product cannibalisation.

5.3 Hawkes Processes for Product Cannibalisation

We are now at the point of defining the model we use to estimate product cannibalisation for wholesale data. This section contains a description of our modelling approach and its intensity function utilising concepts from Chapter 2 and Chapter 4. In addition, we give details on the choice of the background rate $\mu_m(\cdot)$ and the influence kernel $g_{jm}(\cdot)$.

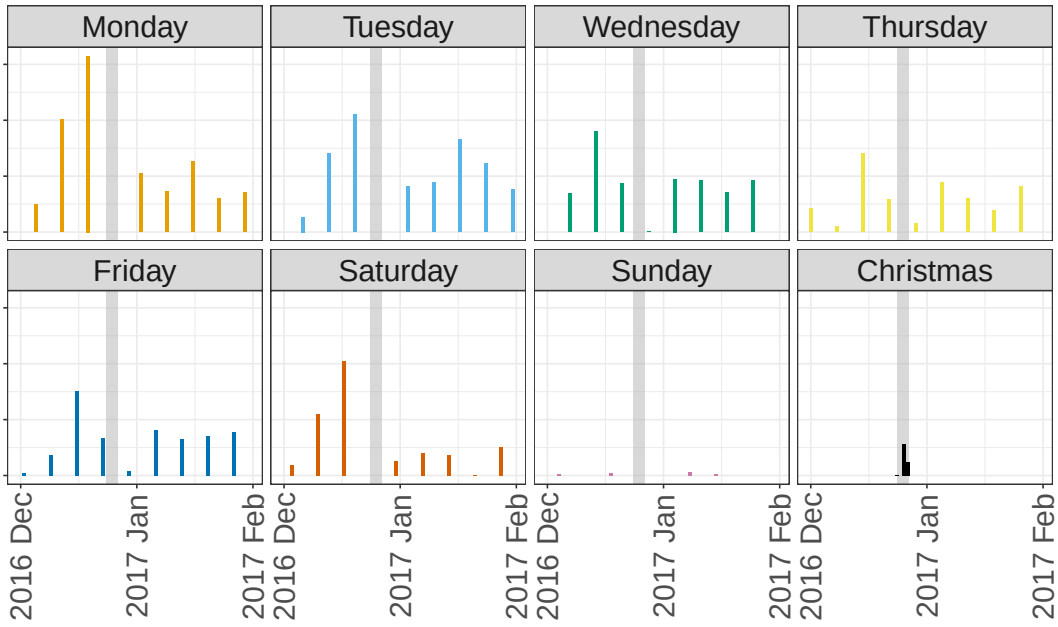


Figure 5.3: Total number of orders placed for products in Product Class B by BusinessGroup for December 2016 and January 2017 per day of the week and Christmas period. The shaded area highlights the Christmas period in each subplot. The y-axis has been removed for data protection reasons.

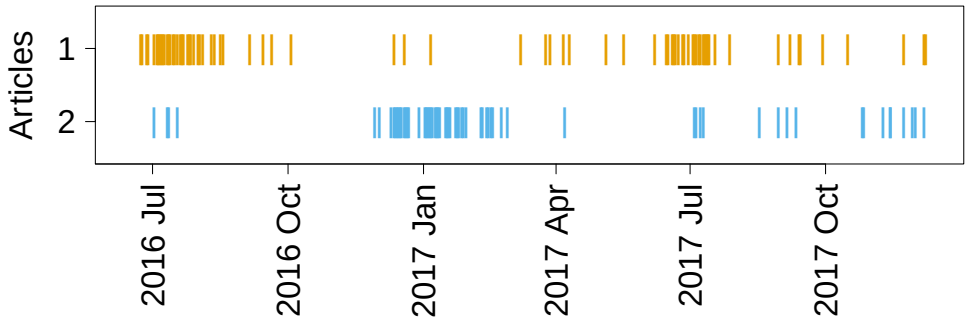


Figure 5.4: Orders placed for two products by BusinessGroup. Each vertical bar indicates that on the particular day an order was placed that included the respective article.

5.3.1 Model

To estimate product cannibalisation we use the multivariate Hawkes process such that each article $m = 1 \dots M$ is represented by a dimension $m = 1 \dots M$. Whenever an order for article m is placed we record an event in dimension m. Based on this data we use the following intensity function for our model

$$\lambda_m(t) = \left[\mu_m(t) + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i = j, \\ t_i < t}} K_{jm}\, g_{jm}(t - t_i) \right]_+ , \tag{5.1}$$

where $[\cdot]_+$ indicates that any negative evaluation inside the brackets is set to zero, as discussed above in our choice of link function.

We restrict the influence such that $K_{jm} < 1$ (hence a negative K is permitted), which means that both excitation and inhibition are allowed in this model. If the interaction K_{jm} for $j \neq m$ is negative we interpret this inhibition as product cannibalisation as the occurrence of an event in dimension j (article j is bought) makes it less likely that an event happens in dimension m (article m is ordered). This is precisely in line with our understanding of product cannibalisation where the purchase of article j makes it less likely that article m is bought. Hence, when K_{jm} is negative, article j cannibalises article m. With that, we have now got a model that can account for product cannibalisation.

Two parts of the intensity function in Equation 5.1 still need to be defined, $\mu_m(\cdot)$ and $g_{jm}(\cdot)$. The subsequent sections provide details on both the background rate and the influence kernel to conclude the definition of our model.

5.3.2 Background Rate

The choice of background rate $\mu_m(\cdot)$ for Hawkes processes is often flexible and application-specific. For example, Mohler (2013) uses a Log-Gaussian Cox process, Molkenthin *et al.* (2022) utilise a Gaussian Process to represent the background rate, and Markwick (2020) employs a Dirichlet process. As the data described in Section 5.2 displays distinct seasonal variability this needs to be taken into account in our estimation procedure through an adapted background rate. We choose for $m = 1 \dots M$

$$\mu_m(t) = c_m\, b(t), \tag{5.2}$$

where $b(\cdot)$ is accounting for general seasonality. The article-specific scaling parameter c_m is flexible enough while remaining computationally cheap. We use the following parametric form for the seasonal part of the background rate:

$$b(t) = \begin{cases} \left[\varphi_1 \mathbb{1}_{\text{Mon}}(t) + \dots + \varphi_7 \mathbb{1}_{\text{Sun}}(t)\right] \left[\varphi_8 \mathbb{1}_{\text{Jan}}(t) + \dots + \varphi_{19} \mathbb{1}_{\text{Dec}}(t)\right], & \text{if } \mathbb{1}_{\text{Christmas}}(t) = 0 \\ \varphi_{20}, & \text{if } \mathbb{1}_{\text{Christmas}}(t) = 1. \end{cases} \tag{5.3}$$

This describes a multiplicative effect between day of the week ($\varphi_1 \dots \varphi_7$) with the month ($\varphi_8 \dots \varphi_{19}$) outside of the Christmas period (24^{th} till 27^{th} of December),

and a constant rate φ_{20} during the Christmas period.

5.3.3 Influence Kernel

The options for influence kernels are equally broad. In some instances, they are highly problem-specific, for example, Browning *et al.* (2021) use a histogram kernel for Covid modelling. A popular choice of the influence kernel is the exponential kernel, which performs well in many examples (e.g. Blundell *et al.*, 2012; Shelton *et al.*, 2018). For the influence kernels, we utilise this exponential kernel

$$g_{jm}(z) = \beta_{jm} \exp\left(-\beta_{jm}\, z\right), \tag{5.4}$$

for $z > 0$ with $\beta_{jm} > 0$ for $j, m = 1 \dots M$. Additionally, we assume that all $\beta_{jj} = \beta_{\text{diag}}$ and $\beta_{jm} = \beta_{\text{off}}$ when $j \neq m$. Hence, the values for β_{jm} are the same for all self-influences, as well as cross-influences, respectively.

5.4 Estimation

To estimate the parameters of the model from Section 5.3 we use a Bayesian approach incorporating the priors that will be discussed in Section 5.5. We utilise Stan (Stan Development Team, 2019) to obtain posterior samples and analyse the posterior distribution through density plots. We utilise our findings from Chapter 4 to deal with challenges arising from the need to integrate the intensity function when evaluating the likelihood. In particular, we employ the approximation for Λ showcased in Section 4.2.2.

Another aspect to consider is the estimation of the background rate $\mu_i(\cdot)$ for $m = 1 \dots M$. As discussed in Section 5.3.2 the background rate consists of two parts

$$\mu_m(t) = c_m\, b(t), \tag{5.5}$$

where c_m is a product specific scaling parameter and $b(\cdot)$ is captures general, seasonal trends. We choose a multiplicative effect between day of the week $(\varphi_1 \dots \varphi_7)$ with the month $(\varphi_8 \dots \varphi_{19})$ outside of the Christmas period (24^{th} till 27^{th} of December), and a constant rate φ_{20} during the Christmas period.

Estimation of components of the background rate can either be done jointly with all other model parameters (Mohler, 2013; Ross and Kolev, 2022; Molkenthin *et al.*, 2022) or outside of the model (Helmstetter *et al.*, 2006). For our approach, we choose to estimate the product-specific scaling parameter c_m within Stan and provide the seasonal component $b(\cdot)$ as a plug-in estimate which is produced beforehand outside of Stan to reduce the number of parameters having to be estimated. For this, we consider all articles in the respective category, for example, all products in Product Class A for the first example in Section 5.6.1. Since the numbers of orders can differ substantially between products, we subsample within each product such that the same number of events per product is used in the construction of the background rate. This avoids the domination of a few high-order products. Once this subset of orders is constructed, we employ an optimiser in R on this data set to obtain the maximum likelihood estimation for

$(\varphi_1 \dots \varphi_{20})$. These are then used to construct the plug-in estimate $b(\cdot)$, which is passed to Stan.

5.5 Prior Choice

To implement the proposed model from Section 5.3.1 for product cannibalisation in a Bayesian framework we investigate prior choices for $\mu_m(\cdot)$, $\mathbf{K}$, and $\boldsymbol{\theta}_g$ (using an exponential kernel).

5.5.1 Background Rate

For each dimension $m = 1 \dots M$ we use $c_m > 0$ with prior

$$c_m \sim \mathcal{N}(0, 3), \qquad \text{for } m = 1 \dots M, \tag{5.6}$$

that scales the plug-in estimates accordingly as the sale volumes may differ between articles.

5.5.2 Influence Magnitude

Throughout the literature, Hawkes processes are parameterised in terms of $\mathbf{K}$. For example, Browning *et al.* (2021) choose a uniform prior for each of its entries. However, when the dimension M is large the non-negative entries of $\mathbf{K}$ have to be smaller to retain stability in accordance with the criteria outlined in Section 4.3. For example, when all entries of $\mathbf{K}$ are 0.4, a two-dimensional process is stable, whereas a three-dimensional $\mathbf{K}$ has an eigenvalue larger than 1 and hence is not stable. Priors on $\mathbf{K}$ in a Bayesian framework would therefore have to be adapted according to the dimension M. Instead, we suggest to reparametrise the model to circumvent this problem using the total number of offsprings, which does not suffer from this dimension-dependency, and use this notion to place priors.

Reparametrise the Model Using the Total Number of Offsprings

In this section, we formally define the total number of offsprings, which is then used for the prior elicitation. For now, we assume that all entries of $\mathbf{K}$ are non-negative and the background rate is constant. Section 2.2.1 introduces the notion of direct offspring, i.e. an event in dimension m from the process of an event in dimension j. However, when estimating the parameters it can become difficult to distinguish between an event in j triggering a process in dimension m ('$j \to m$') and an event in j triggering a process in different dimension k, which produces an event that, in turn, triggers an process in m ('$j \to k \to m$'). This problem is perpetuated further in higher dimensions (Eichler, 2013). To circumvent this issue we propose to investigate the total number of offsprings. For this, two kinds of events are defined as an indirect offspring of the i^{th} event at time t_i:

1. an event from a process that was triggered by a direct offspring of t_i
2. an event from a process that was triggered by an indirect offspring of t_i

We define K^*_{jm} as the total number of offsprings an event in dimension j has in dimension m, which is calculated as the sum of direct and indirect offsprings in dimension m. We write $\mathbf{K}^* = \{K^*_{jm}\}$ were $j, m = 1 \dots M$. In addition, we use $\mathbf{I}$ for the identity matrix of appropriate dimension and $\mathbf{A}^\intercal$ to indicate the transpose of matrix $\mathbf{A}$. This is distinct from T, which is the upper bound for observations. Let us now provide the computation that relates $\mathbf{K}^*$ to $\mathbf{K}$.

Theorem 6. *The expected total number of offsprings is*

$$\mathbf{K}^* = (\mathbf{I} - \mathbf{K})^{-1} - \mathbf{I}.$$

To prove this we first note that $\mathbb{E}[N] = (\mathbb{E}[N_1] \dots \mathbb{E}[N_M])^\intercal$, the total number of expected events in each dimension, can be calculated by $\mathbb{E}[N] = (I - \mathbf{K}^\intercal)^{-1}\boldsymbol{\mu}$, as described by Hawkes (1971); Jovanović *et al.* (2015). In addition, we need an auxiliary Lemma

Lemma 7. *The number of expected events in each dimension is*

$$\mathbb{E}[N] = (\mathbb{E}[N_1] \dots \mathbb{E}[N_M])^\intercal = (\mathbf{K}^* + \mathbf{I})^\intercal \boldsymbol{\mu}.$$

Proof. The Hawkes process can be written as the superposition of Poisson processes as outlined by the branching structure interpretation (see Section 2.2.1). Hence, each N_i consist of two parts:

1. The number of immigrants events from the background process with rate μ_i
2. The number of offsprings in i from an immigrant event in each dimension j

Therefore, we can write

$$\mathbb{E}[N_m] = \mu_m T + \sum_{j=1}^{M} \mu_j T K^*_{jm}. \tag{5.7}$$

In matrix notation this is

$$\mathbb{E}[N] = (\mathbb{E}[N_1] \dots \mathbb{E}[N_M])^\intercal = (\mathbf{K}^* + \mathbf{I})^\intercal \boldsymbol{\mu}. \tag{5.8}$$

□

With that, we can now prove Theorem 6.

Proof. We use the definition of $\mathbb{E}[N]$ from the literature

$$\mathbb{E}[N] = (\mathbf{I} - \mathbf{K}^\intercal)^{-1}\boldsymbol{\mu} \tag{5.9}$$

and relate it to Lemma 7 in the following manner:

$$(\mathbf{I} - \mathbf{K}^\intercal)^{-1}\boldsymbol{\mu} = (\mathbf{K}^* + \mathbf{I})^\intercal \boldsymbol{\mu}. \tag{5.10}$$

Rearranging, this allows us to write $\mathbf{K}^*$ as

$$\mathbf{K}^* = (I - \mathbf{K})^{-1} - \mathbf{I}. \tag{5.11}$$

□

It is important to note that the idea and definition of $\mathbf{K}^*$ have already been presented in the literature, but not in a unified concept towards use in application. On one hand, Bacry and Muzy (2016) introduce this formula without its interpretation in the context of a matrix convolution. On the other hand, Bacry *et al.* (2016) define the concept of total offsprings but do not provide a closed-form expression. Crucially, neither of them make extensive use of the concept, in particular not for prior elicitation.

If a process is stable (see Section 4.3), the calculation remains the same when entries of $\mathbf{K}$ are negative and $\mathbf{K}^*$ still provides meaningful interpretation. While positive entries of $\mathbf{K}^*$ describe the average number of total offsprings, a negative entry summarises the *negative contributions* to the intensity function across dimensions. The number of actually inhibited events depends on the number of events in the process. Nevertheless, $\mathbf{K}^*$ retains its attractive interpretation and can be used to place priors without having to consider the dimensions M.

Crucially, the entries of $\mathbf{K}^*$ for a stable process are not dependent on the dimension M. We therefore reparameterise the multivariate Hawkes process in terms of $\mathbf{K}^*$ such that the intensity for dimension m is

$$\lambda_m(t) = \left[\mu_m(t) + \sum_{j=1}^{M} \sum_{\substack{i:\, d_i = j, \\ t_i < t}} f(\mathbf{K}^*)_{jm}\, g_{jm}(t - t_i) \right]_+, \tag{5.12}$$

where $f(\mathbf{X}) = \mathbf{I} - (\mathbf{X} - \mathbf{I})^{-1}$.

Normal Priors

This reparameterisation in Equation 5.12 permits us to use priors directly for $\mathbf{K}^*$ as these values do not depend on the dimension M. We restrict the parameter space of $\mathbf{K}^*$ such that only stable parameters (according to **C3** from Section 4.3) are allowed. We place independent normal priors on each entry of $\mathbf{K}^*$, such that

$$K^*_{jm} \sim \mathcal{N}(0, 0.5), \qquad \text{for } j, m = 1 \dots M \text{ (stable only)}. \tag{5.13}$$

Note that we do not enforce symmetry or any other structure in $\mathbf{K}^*$.

5.5.3 Influence Kernel

For the influence kernels, we utilise the popular exponential kernel. As discussed in Section 5.3.3, we assume that all $\beta_{jj} = \beta_{\text{diag}}$ and $\beta_{jm} = \beta_{\text{off}}$ when $j \neq m$. Hence, the values for β are the same for all self-influences, as well as cross-influences,

respectively. For β_{diag} and β_{off} we choose the following priors:

$$\beta_{\text{diag}}, \sim \mathcal{U}(0.05, 0.5), \tag{5.14}$$
$$\beta_{\text{off}}, \sim \mathcal{U}(0.05, 0.5). \tag{5.15}$$

As the data is measured in days, the lower bound of the prior ensures that the influence of an event in dimension j onto dimension m is not too far in the future. For example for $\beta_{\text{diag}} = 0.1$, the median of the exponential distribution is approximately 7, which means that half of the influence of an event happens within a week of it. The upper bound of the distribution makes sure that the influence kernel is at least somewhat spread out and not concentrated immediately after an event.

5.6 Application

This section studies two examples of a multivariate Hawkes process with excitation and inhibition to detect product cannibalisation using the model from Section 5.3.1 and the priors from Section 5.5. The first example models two products from Product Class A, whereas the second one examines how the orders of four products in a different Product Class B interact. For data privacy reasons we cannot disclose the nature of these product classes. For each example, we also fit two models without inhibition that serve as comparisons both on the training and test set.

5.6.1 Product Class A

For our first example, we select two similar products from Product Class A to examine the product cannibalisation between them with $M = 2$. Both are similar in their appearance and target audience. Their suggested retail prices differ by approximately 20%.

We use one year (2016-06-14 to 2017-06-13) as training data (a total of 109 observations, 55 for Article 1, 54 for Article 2). The following half year (2017-06-14 to 2017-12-13) is used as a test period (91 events, of which 60 from Article 1 and 31 from Article 2). Figure 5.5 displays these events. As discussed in Section 5.2, we are dealing with wholesale data, hence all sales are to the same wholesale customer (BusinessGroup).

As described in Section 5.4, the plug-in estimate for $b(\cdot)$ of the background rate is based on all articles in the same product class. This ensures that $b(\cdot)$ only captures large, seasonal trends. Given this plug-in estimate, we then use the prior set-up from Section 5.5 to obtain posterior distributions for $\boldsymbol{\theta}$ using Stan (Stan Development Team, 2019). The estimation is carried out on the training set using normal priors that permit inhibition for $\mathbf{K}^*$ (Model 1). Obtaining 750 samples (after 50% burn-in) from three parallel chains takes about 3 hours on a classic laptop. For comparison, we also fit two additional models without inhibition that serve as benchmarks. Model 2 does not allow any inhibition ($0 \leq K_{jm} < 1$ for all j, m), whereas Model 3 only uses the background rate, which means that $K_{jm} = 0$

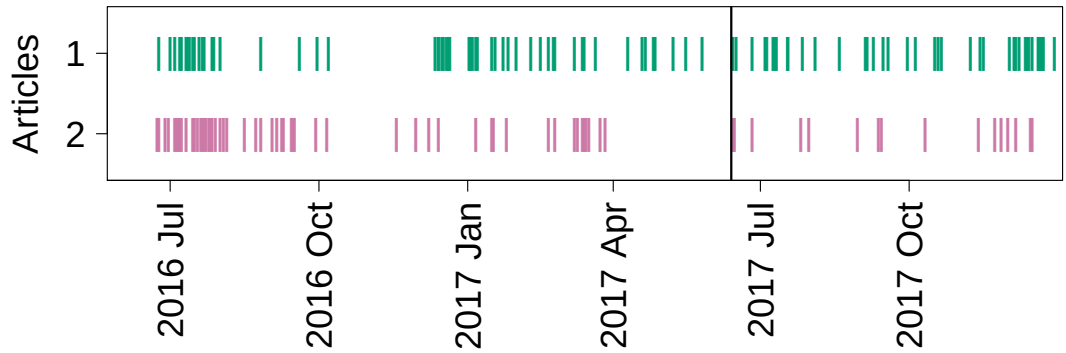


Figure 5.5: Orders placed for two products from product class A by BusinessGroup. Each vertical bar indicates that on the particular day an order was placed that included the respective article. The vertical black line indicates the split between training and test data.

for all j, m. Figure 5.6 showcases the posterior distribution for $\mathbf{K}^*$ for Model 1 and Model 2. Model 3 is not included in the plot as its posterior distribution is simply a point mass at zero for each entry.

For the model that allows inhibition (Model 1) the parameter K_{12}^* is estimated to be negative. We estimate $K_{12}^* = -0.49$ with a 90%- credible interval of $(-0.90, -0.10)$. We can interpret this as product cannibalisation in the sense that Article 1 cannibalises sales of Article 2. Interestingly, the other cross-influence parameter K_{21}^* is estimated as (close to) zero. As the two articles are very similar from an appearance perspective, their main difference lies in the suggested retail price. This analysis suggests that the wholesale customer is buying the slightly cheaper article instead of the more expensive one, but not vice versa.

In addition, we can also examine how long the self and cross-influence last. The posterior means for in Model 1 are $\beta_{\text{diag}} = 0.14$ $(0.11, 0.49)$ and $\beta_{\text{off}} = 0.33$ $(0.11, 0.49)$. This leads to to the influence kernels (for $K = 1$) as showcased in Figure 5.7 for the self and cross-influence. Half of the self-influence takes place in the first five days after the event, whereas this number sits at two days for the cross-influence. This means that the cross-inhibitory effects are most pronounced immediately after an order is placed.

Table 5.2 compares our suggested model incorporating inhibition (Model 1) to two benchmark models which do not allow inhibition. This comparison is done both within the training set (training set log-likelihood and DIC), as well as the predictive likelihood on the test set. Across all comparisons, the inhibition-encompassing Model 1 shows the best performance. Both in the training and test set the models without inhibition (Model 2 and Model 3) give worse outcomes. These results clearly show the need for inhibition and hence product cannibalisation when modelling the sales process using point processes.

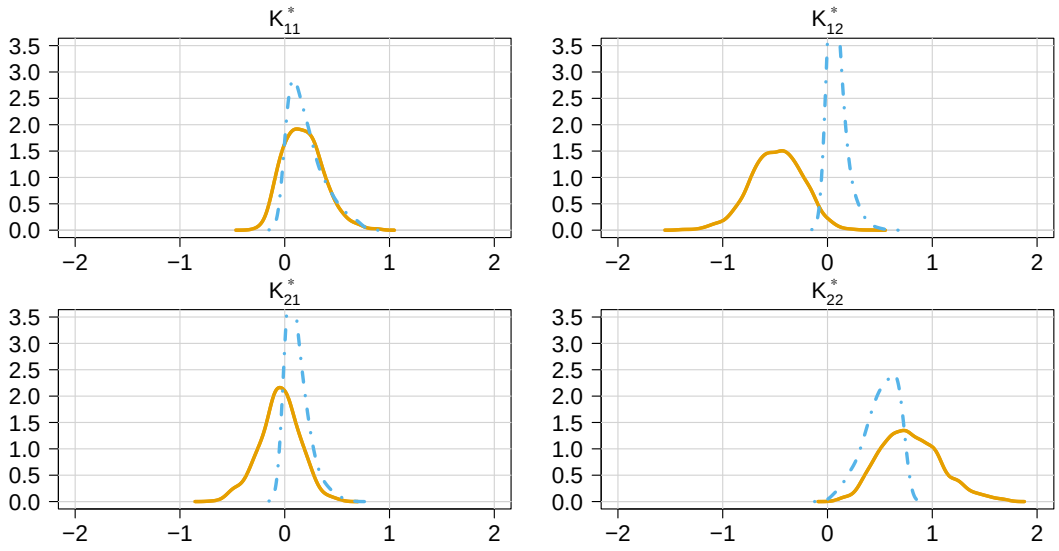


Figure 5.6: Posterior density estimates for the entries of $\mathbf{K}^*$ based on orders placed for two products Business on the training set. The orange solid line represents the normal priors (Model 1), and the blue dot-dashed line is the excitation-only reference model (Model 2).

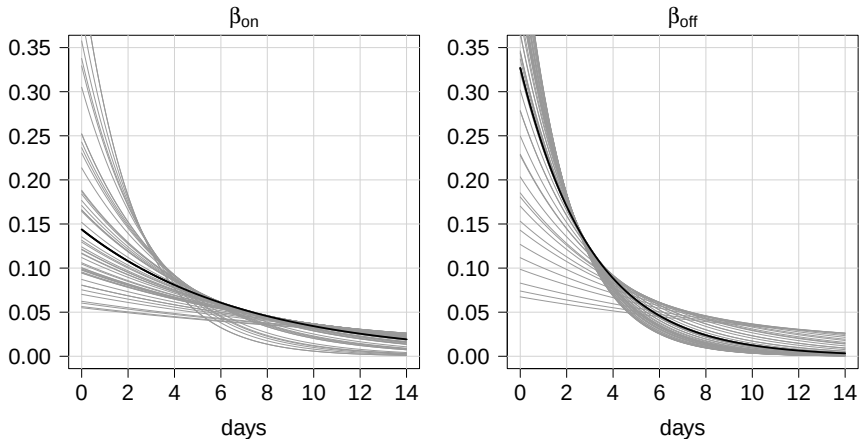


Figure 5.7: Plots of the influence kernels (self and cross-influence) based on the estimated of Model 1 on the training set for $M = 2$ articles. The black line shows the influence kernel at the posterior mean, the grey lines show the influence kernels from 50 posterior samples.

Table 5.2: Comparisons of three models for $M = 2$.

	Model	**Restriction**	**loglik train**	**DIC**	**predictive loglik**
1	Normal Prior	$K_{j,m} < 1$	**-273.39**	**557.44**	**-169.12**
2	Excitation only	$0 \leq K_{jm} < 1$	-278.14	562.36	-180.18
3	Background only	$K_{jm} = 0$	-282.27	568.38	-172.34

Table 5.3: Information on four articles used in the example in Section 5.6.2. Columns 2-4 describe the articles. The last column contains the number of orders placed by BusinessGroup from 2017-06-14 to 2017-12-13 for this article in the training (test) period.

Article	Appearance	Details	Label	Price	Orders train (test)
1	dark	white	none	low	63 (11)
2	light	colour	SomeLabel	high	148 (94)
3	dark	minimal	SomeLabel	high	30 (20)
4	light	minimal	SomeLabel	medium	105 (7)

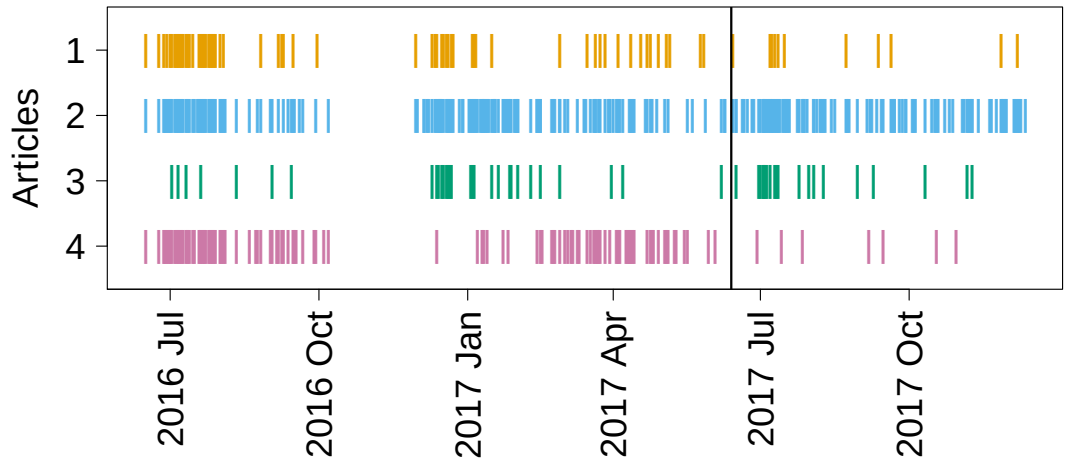


Figure 5.8: Orders placed for products in Product Class B by BusinessGroup. Each vertical bar indicates that on the particular day an order was placed that included the respective article. The vertical black line indicates the split between training and test data.

5.6.2 Product Class B

In our second example, we look at the orders placed by BusinessGroup for four similar products from Product Class B, $M = 4$. Table 5.3 gives some characteristics of these articles. We use one year as a training period (2016-06-14 to 2017-06-13) and the consecutive half year as a test period (2017-06-14 to 2017-12-13). The occurrences of events are displayed in Figure 5.8 and Table 5.3 gives the number of orders per article in the train and test set.

For the plug-in estimate for $b(\cdot)$ of the background rate, we use all products in the same class that have sales in the relevant period. As in the above Section 5.6.1 we use the prior set-up from Section 5.5 to obtain posterior distribution samples for $\boldsymbol{\theta}$ using Stan (Stan Development Team, 2019). For our proposed model, generating 750 samples (after 50% burn-in) from three parallel chains takes about 11 hours on a classic laptop.

As outlined above, we fit one proposed model that incorporates inhibition and two benchmark models without inhibition. Model 1 uses normal priors which per-

Table 5.4: Comparisons of three models for $M = 4$.

	Model	Restriction	loglik train	DIC	predictive loglik
1	Normal Prior	$K_{jm} < 1$	**-678.73**	**1381.45**	**-297.34**
2	Excitation only	$0 \leq K_{jm} < 1$	-698.10	1411.39	-313.32
3	Background only	$K_{jm} = 0$	-726.33	1460.49	-349.23

mit inhibition, whereas Model 2 is excitation-only and Model 3 is only modelled by the scaled background rate. Figure 5.9 plots the obtained posterior distributions of the entries of $\mathbf{K}^*$ for Model 1 and Model 2. For Model 1, two parameters indicate product cannibalisation. We estimate $K^*_{32} = -0.60$ $(-1.15, 0.00)$ and $K^*_{43} = -0.40$ $(-0.82, 0.00)$. We conclude that orders for Article 3 cannibalise orders of Article 2, as they are both at a higher price point and part of SomeLabel. In addition, orders for Article 4 cannibalise orders for Article 3. As above, we see that the cheaper article (regarding the suggested retail price) cannibalises the more expensive one, but not vice versa. Also note that Article 1 is not affected by any inhibition, potentially due to the lower price point, the distinct design (white details), and the lack of label affiliation, all of which seem to render it unsuitable to potentially substitute the other articles.

We also examine the shape of the influence kernels (both self and cross-influence) in Figure 5.10. The self-influence kernel has a median of nine days and the cross-influence kernel's median lies at two days. This suggests again that any cross-influences, such as inhibition, are most influential in the days immediately after an order was placed. While this plot shows rather low uncertainty in the kernel, this exponential kernel is then multiplied by an entry of $\mathbf{K}$, which carries significant variance in line with Figure 5.9. The overall influence uncertainty is then driven by uncertainty in both magnitude and kernel.

Table 5.4 compares the three fitted models both on the training set (using the log-likelihood as well as the DIC) and the test set (predictive log-likelihood). Model 1 with inhibition achieves the highest log-likelihood and lowest DIC in the training set, as well as the highest predictive likelihood on the test data. As above, models without inhibition (Model 2 and Model 3) are distinctively sub-par across all metrics. This clearly warrants the consideration of inhibition and hence product cannibalisation when modelling the orders for similar articles.

5.7 Discussion

In this chapter, we focused on estimating product cannibalisation for wholesale data using the multivariate Hawkes process. We used our proposed model to estimate product cannibalisation for $M = 2$ and $M = 4$ articles and compared our suggested model incorporating inhibition to two reference models without inhibition. The superior performance of the models with inhibition across the board gives a strong mandate to consider product cannibalisation when modelling wholesale orders.

We would like to extend this work to higher dimensions and incorporate additional covariates, as our data is rich in articles and information about them. This

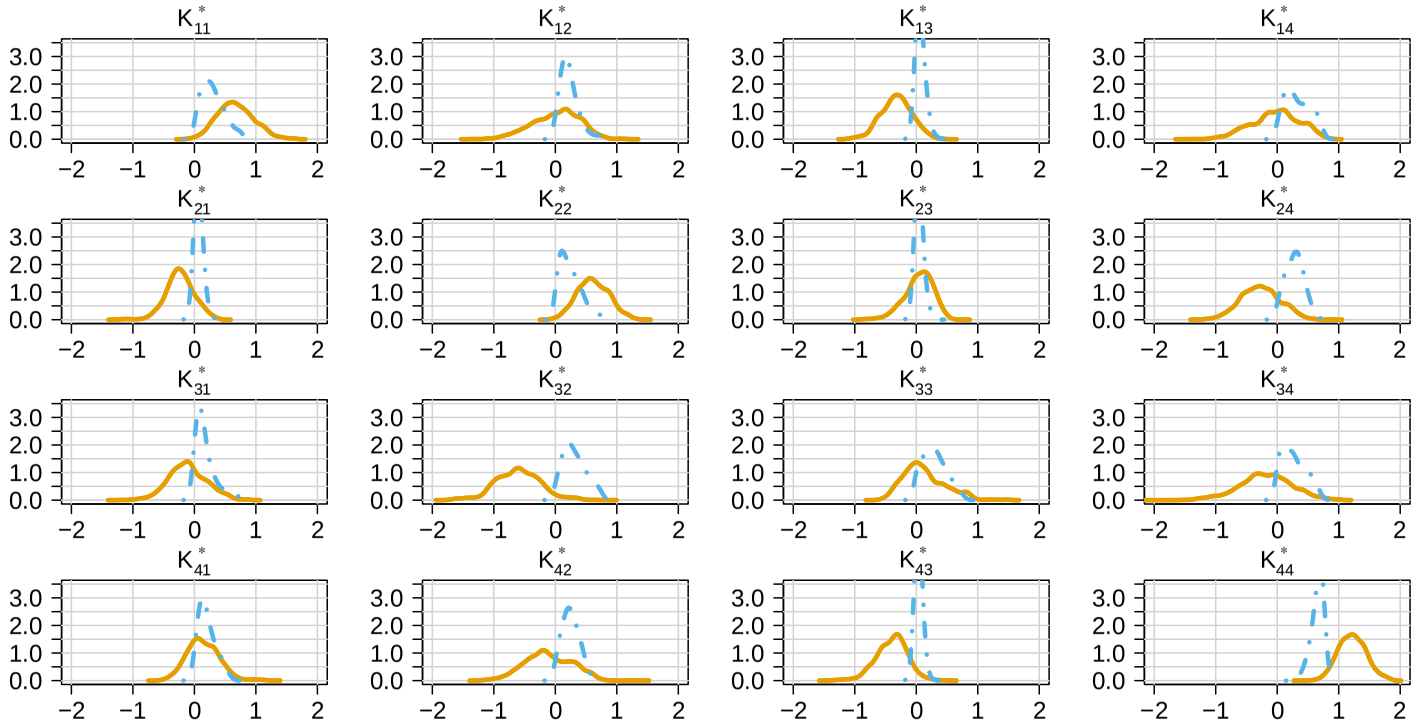


Figure 5.9: Posterior density estimates for the entries of $\mathbf{K}^*$ based on orders placed for products in Product Class B by BusinessGroup on the training set. The orange solid line represents the normal priors (Model 1), and the blue dot-dashed line is the excitation-only reference model (Model 2).

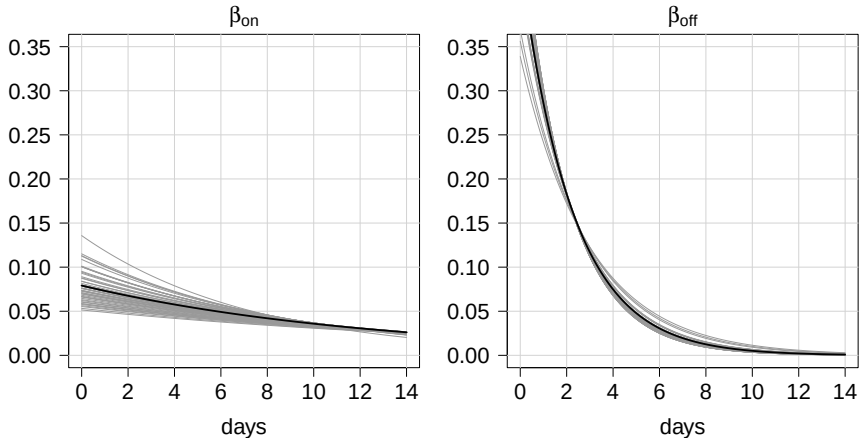


Figure 5.10: Plots of the influence kernels (self and cross-influence) based on the estimated of Model 1 on the training set for $M = 4$ articles. The black line shows the influence kernel at the posterior mean, the grey lines show the influence kernels from 50 posterior samples.

would warrant both computational considerations (e.g. introduce an upper limit for the influence to ease the likelihood calculation), as well as structural ones (e.g. regularisation). A logical next step is the inclusion of suggested retail price or order size, which parallels the idea of marked point processes in the earthquake literature (see, for example, Schoenberg, 2003), that lay beyond the scope of this chapter. As the data is recorded in days, we would also like to explore whether a multivariate Hawkes process in discrete time (as used by Browning *et al.*, 2021) could bring additional benefits.

Finally, this work represents a new way of estimating product cannibalisation, in particular from the wholesale perspective. We therefore aim to apply our method to different scenarios from a variety of industries to explore where business insights can be generated.

Chapter 6

Ancestor Hawkes Model

In this chapter, we propose the Ancestor Hawkes model, which incorporates the latent branching structure (see Section 2.2.1 and Section 2.2.2) into the parameter structure. This permits immigrant and triggered events to have different influences. The No-Cascade Hawkes is introduced as a restricted variant, where the triggering influence magnitude of non-immigrant events is restricted further. We also provide a computationally efficient sampler for the posterior distributions. This is then showcased on real-world group chat data. To the best of our knowledge, this is one of the first examples where group chat data is examined.

Section 6.1 motivates both the Ancestor Hawkes and No-Cascade models. These models are then formally defined in Section 6.2. Section 6.3 provides details on estimation procedures using Gibbs samplers for the proposed models, as well as the classic multivariate Hawkes process. A simulated data example is given in Section 6.4. Section 6.5 contains the example on real data from a group chat setting. We finish with a discussion of our work in Section 6.6.

6.1 Motivation

The classic Hawkes process assumes that every event has the same influence magnitude, irrespective of the branching structure (introduced in Sections 2.2.1 and 2.2.2). However, in some applications, it might be reasonable for immigrant and triggered events to have different influence magnitudes. Therefore, we suggest to control the influence of an event based on the fact that they are immigrant or triggered events. To that end, we propose two new Hawkes process models, the Ancestor Hawkes and the No-Cascade Hawkes. Note that, despite the similarity in name, the proposed Ancestor Hawkes approach is not related to Ancestor sampling (Lindsten *et al.*, 2012).

The first model is the broader *Ancestor Hawkes* model. Its goal is to permit immigrant and triggered events to have different influences. Instead of just one matrix $\mathbf{K}$ of influence magnitudes for all events as in the classic Hawkes, two such matrices are considered. Now $\mathbf{K}$ governs the influence magnitude immigrant events have on all dimensions and $\mathbf{L}$ describes the influence magnitude that triggered events have.

As a special case, we propose the *No-Cascade* Hawkes approach. It is designed

to stop cascades running through multiple dimensions. Here, only immigrant events can trigger events in dimensions that are not their own. Both immigrant and triggered events can trigger in their own dimensions. This prohibits cascades from running across more than two dimensions and clearly separates direct ($\mathbf{K}$) and indirect ($\mathbf{L}$) influence magnitudes.

Both models are a substantial extension of models currently utilised in the literature. Li *et al.* (2020) estimate a branching structure for Hawkes processes where the branching structure is not available (e.g. inhibition, non-linear link function). They then use an adapted Chinese Restaurant Process to allow different triggerings per cluster of events. Kolev and Ross (2019) make a distinction between immigrant and triggered events in a Hawkes process model. They apply their approach to earthquakes where the arrival of a main shock depends on the time passed since the last main shock. This allows for different event arrival dynamics for immigrant and triggered events but does not assign different influences based on the branching structure.

Let us motivate the Ancestor Hawkes model in an example. A prime application for it is the group chat setting, which we investigate in Section 6.5. A group chat is a synchronised online conversation between more than two participants. We focus on private group chats where only admitted participants can see the conversation and every participant has an equal opportunity to send messages (Mannell, 2020).

The literature on group chat modelling is sparse. For one of the few examples see Guo *et al.* (2019), who use a long short-term memory neural network. Classic Hawkes processes have been used to model one-to-one conversations such as emails (Miscouridou *et al.*, 2018), as well as one-to-many scenarios, like tweets on Twitter/X (Rizoiu *et al.*, 2017). However, they have not been used for small-scale group chats, where all participants contribute actively.

We propose the Ancestor Hawkes model for group chats, as it can capture dynamics in a group chat beyond the classic Hawkes process. The set-up is as follows. Each of the chat participants is represented by a dimension. When they send a message, an event is recorded in that dimension. Consider three people A, B, and C in a group chat for a toy example. We now examine the reply rates of Person C in two different scenarios with the following messages being sent.

1. Person A: *"Who would like to come to the pub tonight?"* (immigrant event)

2. Person B: *"Who would like to come to the pub tonight?"* (immigrant event)
   Person A: *"I would love to come!"* (triggered event)

In the classic Hawkes model, the message from A in both the first and the second scenario increases the intensity of Person C according to $K_{A\to C}$. This implies that Person C has the same rate of replies to messages *"Who would like to come to the pub tonight?"* and *"I would love to come!"* as both messages come from Person A. Since they are different in the context of the conversation (immigrant event vs. triggered reply), this might not be appropriate for this application.

The Ancestor Hawkes model alleviates this. It allows for different reply rates for messages that start a new messaging cascade (immigrant events) and for those

that are a contribution/reply to an ongoing conversation (triggered event). Here, $\mathbf{K}$ governs the influence of immigrant events and $\mathbf{L}$ describes the influence of triggered events. Hence, *"Who would like to come to the pub tonight?"* increases the intensity of Person C by $K_{A\to C}$, since this is an immigrant message. In contrast, the message *"I would love to come!"* increases the intensity of Person C according to $L_{A\to C}$, as it is a triggered message.

This allows the rate of answer for $A \to C$ to differ between immigrant ($K_{A\to C}$) and triggered events ($L_{A\to C}$). We show below in Section 6.5 that through this the Ancestor Hawkes model can capture dynamics in the group chat data that the classic Hawkes is not able to express.

Our other proposed approach goes one step further. The No-Cascade Hawkes disallows any cascades for from running across more than two dimensions. In the example above, this implies that the increase in intensity of Person C in response to the triggered message *"I would love to come!"* is zero. While this is a strong restriction of the model, it has the attractive advantage of making the parameter interpretation easier and hence might be appropriate in some applications. In Chapter 5 we introduced $\mathbf{K}^*$, which summarised all direct and indirect influences of an event across dimensions. The No-Cascade Hawkes instead limits the cascading behaviour, such that $\mathbf{K}$ contains all direct influences and $\mathbf{L}$ contains all indirect ones. This simplifies otherwise difficult-to-interpret influences in high-dimensional problems, as discussed by Eichler (2013).

6.2 Models

This section introduces additional notation needed to describe the two proposed models. Subsequently, the classic Hawkes process is reviewed and the new models are introduced. Note that the No-Cascade Hawkes process is a restricted version of the more general Ancestor Hawkes. For ease of exposition, a constant background rate is assumed. Extending these models to a non-constant background rate is straightforward.

6.2.1 Notation

The notation for events remains the same as in the previous chapters. As a reminder, let us consider a multivariate Hawkes process in M dimensions on $[0, T]$. There are a total of N events. An event is written as $y_i = (t_i, d_i)$ for $i = 1 \dots N$. In addition, let

$$B_i = \begin{cases} 0, & \text{if } t_i \text{ was produced by the background process (it is an immigrant event)} \\ k, & \text{if } t_i \text{ was triggered by } t_k \text{ (it is a triggered event)} \end{cases} \tag{6.1}$$

and $\mathbf{B} = \{B_1 \dots B_N\}$, which describe the branching structure (see Section 2.2.1 and Section 2.2.2 for details on the branching structur). Further variables for the branching structure based on $\mathbf{B}$ are defined as

$$S_{i,m} = \{j : B_j = i \text{ and } d_j = m) \text{ for } i = 0 \dots N, m = 1 \dots M, \tag{6.2}$$

where $S_{i,m}$ contains the indices of events in dimension m that are triggered by the process of event i. When $i = 0$, they are immigrant events, otherwise, they were triggered by the event of the respective index. We write $\mathbf{S} = \{S_{i,m}\}_{m=1\ldots M}^{i=0\ldots N}$.

6.2.2 Classic Hawkes

First, let us recall the classic Hawkes process model, which was used in the previous chapters. The standard multivariate Hawkes process has intensity function

$$\lambda_m(t) = \mu_m + \sum_{i:t_i<t} K_{d_i \to m}\, g_{d_i \to m}(t - t_i). \tag{6.3}$$

The parameters for this process are $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{K}, \boldsymbol{\theta}_g)$. For this model, the influence of immigrant and triggered events in dimensions j onto dimension m is the same. Hence, there are a total of M^2 different influence magnitudes in this model. Furthermore, $\boldsymbol{\theta}_g$ parameterises the influence kernel. We use $\theta_{g|m_1 \to m_2}$ to refer to the parameters that govern the influence from dimension m_1 onto m_2.

6.2.3 Ancestor Hawkes

Let us now examine our proposed approach, the Ancestor Hawkes model. Here, immigrant and triggered events differ in their influences. Hence, the branching structure is directly incorporated into the parameter architecture. Here, the influence of an event y_i onto dimension m does not only depend on the dimension of y_i but additionally on its branching variable B_i. When $B_i = 0$ and hence y_i is an immigrant event, then it influences other dimensions according to $\mathbf{K}$. Alternatively, if y_i is a triggered event and therefore $B_i > 0$ then the influence is governed by $\mathbf{L}$. For the Ancestor Hawkes model, we propose intensity

$$\lambda_m(t) = \mu_m + \underbrace{\sum_{\substack{i:t_i<t,\\ B_i=0}} K_{d_i \to m}\, g_{d_i \to m}(t - t_i)}_{\text{contributions from immigrant events}} + \underbrace{\sum_{\substack{j:t_j<t,\\ B_j>0}} L_{d_j \to m}\, h_{d_j \to m}(t - t_j)}_{\text{contributions from triggered events}}. \tag{6.4}$$

This leads to parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{K}, \mathbf{L}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_h)$, where both $\mathbf{K}$ and $\mathbf{L}$ are matrices of dimension $M \times M$. Hence, there are $2M^2$ influence magnitude parameters in this model. In addition, $\boldsymbol{\theta}_g$ parameterises the $g_{m_1 \to m_2}(\cdot)$, the influence kernel for immigrant events, and $\boldsymbol{\theta}_h$ contains the parameters for all $h_{m_1 \to m_2}(\cdot)$, the influence kernel for triggered events. When referring to the parameters that govern the influence from dimension m_1 onto m_2 we write $\theta_{h|m_1 \to m_2}$.

6.2.4 No-Cascade Hawkes

In addition, we propose the No-Cascade Hawkes model. Its goal is to stop cascades running through multiple dimensions to clearly distinguish direct and indirect influences. To that end, triggered events can only trigger in their own dimension according to $\mathbf{L}$. Immigrant events, however, can trigger events in any dimension as governed by $\mathbf{K}$. This set-up prevents cascades from traversing

through more than two dimensions, which makes interpretation easier (Eichler, 2013). The intensity of a No-Cascade Hawkes is

$$\lambda_m(t) = \mu_m + \underbrace{\sum_{\substack{i:t_i<t,\\ B_i=0}} K_{d_j\to m}\, g_{d_j\to m}(t-t_i)}_{\text{contributions from immigrant events}} + \underbrace{\sum_{\substack{j:t_j<t,\\ B_j>0,\\ d_j=m}} L_{m\to m}\, h_{m\to m}(t-t_j)}_{\text{contributions from triggered events in } m} \quad . \tag{6.5}$$

The parameters are again $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{K}, \mathbf{L}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_h)$. As above, $\mathbf{K}$ is a $M \times M$ matrix, but now $\mathbf{L}$ is a vector of length M. Hence, there are $M^2 + M$ influence magnitude parameters. The No-Cascade Hawkes can be seen as a restricted version of the Ancestor Hawkes model, where $\mathbf{L}$ has a non-zero diagonal and zeros elsewhere.

6.3 Estimation

This section gives details on the estimation procedure for the classic Hawkes process in multiple dimensions and offers a similar approach for the proposed models. To sample from the posterior distributions of the Ancestor Hawkes model (and hence the No-Cascade) we employ a Gibbs sampling approach based on the branching structure, which induces conditional independencies. This is paralleling the work of Rasmussen (2013) and Ross (2021). For both, we make use of Metropolis-Hastings proposals (Section 2.3.3).

We assume that priors $\pi(\cdot)$ have been chosen for each parameter. The iterative Gibbs procedure uses a starting value $\boldsymbol{\theta}^{(0)}$ and at step k of the procedure we have parameters $\boldsymbol{\theta}^{(k)}$. All parameters are updated at each step. This is repeated until a certain number of samples is reached. A proportion of earlier samples is removed as burn-in and the remaining samples are additionally thinned. The resulting samples are then draws from the posterior distribution of $\boldsymbol{\theta}$ in a typical Gibbs fashion.

6.3.1 Classic Hawkes with Branching Structure

First, we examine the sampler for a classic Hawkes process in multiple dimensions. Note that this is the straightforward extension of Rasmussen (2013); Ross (2021) into multiple dimensions for unmarked Hawkes processes. Here $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\mu}, \mathbf{K}, \boldsymbol{\theta}_g)$.

As described above, the intensity at time t and dimension m is

$$\lambda_m(t) = \mu_m + \sum_{i:t_i<t} K_{d_i\to m}\, g_{d_i\to m}(t-t_i). \tag{6.6}$$

We also assume that we have access to $G_{m_1\to m_2}(\cdot)$, where $G_{m_1\to m_2}(z) = \int_0^z g_{m_1\to m_2}(x)\, dx$.

Conditional Likelihood

Conditional on the branching structure $\mathbf{B}$ the likelihood of a classic multivariate Hawkes process can be written as

$$p(Y \mid \boldsymbol{\theta}, \mathbf{B}) = \prod_{m=1}^{M} \mu_m^{|S_{0,m}|} \exp(-\mu_m T) \times \prod_{j=1}^{M} \prod_{p:d_p=j} \exp(-K_{j\to m} G_{j\to m}(T - t_p)) \times \prod_{i \in S_{p,m}} K_{j\to m} g_{j\to m}(t_i - t_p). \tag{6.7}$$

This shows that, conditional on the branching variables, the parameters for the background rate $\boldsymbol{\mu}$ are independent of $(\mathbf{K}, \boldsymbol{\theta}_g)$, which describe the influences. In addition, $(K_{jm}, \theta_{g|jm})$ are conditionally independent of $(K_{in}, \theta_{g|in})$ when $i \neq j$ or $n \neq m$. These conditional independencies are exploited in the subsequent sampling procedure, which produces less correlated samples than a random walk MCMC approach (Ross, 2021). In addition, these conditional independencies also facilitate better convergence for the maximum likelihood estimator in an Expectation-Maximisation approach (Veen and Schoenberg, 2008).

We now outline step k of the sampling procedure.

Sample $\mathbf{B}^{(k+1)}$

First, the branching structure is sampled. Here, we assume that the prior places equal prior probability on $B_i = 0 \dots i-1$, which leads to

$$p(B_i^{(k+1)} = j \mid \boldsymbol{\mu}^{(k)}, \mathbf{K}^{(k)}, \boldsymbol{\theta}_g^{(k)}) \propto \begin{cases} \mu_{d_i}^{(k)}, & \text{if } j = 0 \\ K_{d_j \to d_i}^{(k)} \, g_{d_j \to d_i}^{(k)}(t_i - t_j), & \text{if } j \in \{1, 2 \dots i-1\}. \end{cases} \tag{6.8}$$

Each $B_i^{(k+1)}$ for $i = 1 \dots N$ is sampled independently. Based on $\mathbf{B}^{(k+1)}$, the construction of $\mathbf{S}^{(k+1)}$ follows suit.

Sample $\boldsymbol{\mu}^{(k+1)}$

The conditional posterior density of μ_m is

$$\pi(\mu_m \mid \mathbf{B}) \propto \pi(\mu_m) \times \mu_m^{|S_{0,m}|} \exp(-\mu_m T). \tag{6.9}$$

When the prior $\pi(\mu_m)$ is chosen to be a Gamma distribution with parameters (a_μ, b_μ), then the posterior distribution is conjugate with

$$\pi(\mu_m \mid \mathbf{B}) = \text{Gamma}(a_\mu + |S_{0,m}|, b_\mu + T). \tag{6.10}$$

This means each $\mu_m^{(k+1)}$ can be sampled directly from $\text{Gamma}(a_\mu + |S_{0,m}^{(k+1)}|, b_\mu + T)$ for $m = 1 \dots M$.

Sample $\mathbf{K}^{(k+1)}$

In total, there are M^2 entries in $\mathbf{K}$. For a given $m_1 = 1 \dots M$ and $m_2 = 1 \dots M$, the conditional posterior distribution of that particular entry is

$$\pi(K_{m_1 \to m_2} \,|\mathbf{B}, \boldsymbol{\theta}_g) \propto \pi(K_{m_1 \to m_2}) \times \prod_{j:d_j = m_1} \exp(-K_{m_1 \to m_2} G_{m_1 \to m_2}(T - t_j)) \, K_{m_1 \to m_2}^{|S_{j,m_2}|}, \tag{6.11}$$

where $G_{m_1 \to m_2}(z) = \int_0^z g_{m_1 \to m_2}(x)\, dx$.

While there is no conjugate form available, we can sample from this in the usual Metropolis-Hastings manner taking $\pi(K_{m_1 \to m_2} \mid \mathbf{B}, \boldsymbol{\theta}_g)$ as its target using some kind of proposal distribution $q(K_{m_1 \to m_2}^{(k+1)} \mid K_{m_1 \to m_2}^{(k)})$. This distribution should be easy to evaluate and easy to sample. Each entry of $\mathbf{K}$ is sampled independently, where $m_1 = 1 \dots M$ and $m_2 = 1 \dots M$.

Sample $\boldsymbol{\theta}_g^{(k+1)}$

The conditional posterior distribution for the parameters of $g_{m_1 \to m_2}(\cdot)$ is

$$\pi(\theta_{g|m_1 \to m_2} \,|\mathbf{B}, \mathbf{K}) \propto \pi(\theta_{g|m_1 \to m_2}) \times \prod_{j:d_j = m_1} \exp(-K_{m_1 \to m_2} G_{m_1 \to m_2}(T - t_j)) \prod_{i \in S_{j,m_2}} g_{m_1 \to m_2}(t_i - t_j). \tag{6.12}$$

As above, this is sampled using a proposal distribution $q(\theta_{g|m_1 \to m_2}^{(k+1)} \mid \theta_{g|m_1 \to m_2}^{(k)})$ for each entry of $\boldsymbol{\theta}_g$ independently.

6.3.2 Ancestor Hawkes with Branching Structure

Now we lay out the sampling procedure for our proposed Ancestor Hawkes model. Since the No-Cascade Hawkes is a restricted version of the Ancestor Hawkes, its sampling procedure is a minor extension and hence is not discussed separately. At time t the intensity of the Ancestor Hawkes in dimension m is

$$\lambda_m(t) = \mu_m + \sum_{i:t_i < t, B_i = 0} K_{d_i \to m}\, g_{d_i \to m}(t - t_i) + \sum_{j:t_j < t, B_j > 0} L_{d_j \to m}\, h_{d_j \to m}(t - t_j). \tag{6.13}$$

There are now more parameters in the model: $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\mu}, \mathbf{K}, \mathbf{L}, \boldsymbol{\theta}_g, \boldsymbol{\theta}_h)$, where $\boldsymbol{\theta}_g$ parameterises all $g_{m_1 \to m_2}(\cdot)$ and $\boldsymbol{\theta}_h$ provides the parameters for all $h_{m_1 \to m_2}(\cdot)$. Here, $\mathbf{K}$ describes the influence of immigrant events and $\mathbf{L}$ the influence of triggered (= non-immigrant) events. As above, we assume that we have access to the $G_{m_1 \to m_2}(z) = \int_0^z g_{m_1 \to m_2}(x)\, dx$ and $H_{m_1 \to m_2}(z) = \int_0^z h_{m_1 \to m_2}(x)\, dx$.

Conditional Likelihood

The likelihood for data Y under the Ancestor Hawkes model conditional on the branching structure is

$$p(Y \mid \boldsymbol{\theta}, \mathbf{B}) = \prod_{m=1}^{M} \mu_m^{|S_{0,m}|} \exp(-\mu_m T) \times \tag{6.14}$$

$$\prod_{j=1}^{M} \prod_{\substack{p:d_p=j,\\ B_p=0}} \exp(-K_{j\to m}\, G_{j\to m}(T - t_p)) \prod_{i \in S_{p,m}} K_{j\to m} g_{j\to m}(t_i - t_p) \times$$

$$\prod_{\substack{q:d_q=l,\\ B_q>0}} \exp(-L_{l\to m}\, H_{l\to m}(T - t_q)) \prod_{r \in S_{q,m}} L_{l\to m} h_{l\to m}(t_r - t_q). \tag{6.15}$$

As above, this reveals useful conditional independencies that can be exploited in the estimation procedure. Here, $\boldsymbol{\mu}$, as well as $(\mathbf{K}, \boldsymbol{\theta}_g)$, and $(\mathbf{L}, \boldsymbol{\theta}_h)$ are all independent of each other, given $\mathbf{B}$. In addition, $(K_{jm}, \theta_{g|jm})$ are conditionally independet of $(K_{in}, \theta_{g|in})$ when $i \neq j$ or $n \neq m$. The same holds for $(L_{jm}, \theta_{h|jm})$ and $(L_{in}, \theta_{h|in})$. As above, this allows us to design an estimation procedure that improves on a random-walk MCMC, which would suffer from high autocorrelation and hence a lower effective sample size (Ross, 2021). These independencies could also be used for a maximum likelihood estimation in line with Veen and Schoenberg (2008).

With assumed prior distributions $\pi(\cdot)$ for all parameters we now outline step k of the sampling procedure.

Sample $\mathbf{B}^{(k+1)}$

First, the branching structure is sampled assuming that there is an equal prior probability of $B_i = 0 \dots i-1$. Unlike above, the sampling needs to be done sequentially, as the B_i depends on $(B_1 \dots B_{i-1})$ as the contributions per event differ for immigrant and triggered events. This leads to the following sampling procedure:

$$p(B_i^{(k+1)} = j \mid \boldsymbol{\mu}^{(k)}, \mathbf{K}^{(k)}, \mathbf{L}^{(k)}, \boldsymbol{\theta}_g^{(k)}, \boldsymbol{\theta}_h^{(k)}) \propto \tag{6.16}$$

$$\propto \begin{cases} \mu_{d_i}^{(k)}, & \text{if } j = 0 \\ K_{d_j\to d_i}^{(k)}\, g_{d_j\to d_i}^{(k)}(t_i - t_j), & \text{if } j \in \{1, 2 \dots i-1\} \text{ and } B_j^{(k+1)} = 0 \\ L_{d_j\to d_i}^{(k)}\, h_{d_j\to d_i}^{(k)}(t_i - t_j), & \text{if } j \in \{1, 2 \dots i-1\} \text{ and } B_j^{(k+1)} > 0. \end{cases}$$

Based on $\mathbf{B}^{(k+1)}$, the construction of $\mathbf{S}^{(k+1)}$ follows suit.

Sample $\boldsymbol{\mu}^{(k+1)}$

The posterior density of μ_m is the same as in the classic Hawkes case, namely

$$\pi(\mu_m \mid \mathbf{B}) \propto \pi(\mu_m) \times \mu_m^{|S_{0,m}|} \exp(\mu_m T). \tag{6.17}$$

The sampling procedure is identical to the one described above.

Sample $\mathbf{K}^{(k+1)}$

In total, there are M^2 entries in $\mathbf{K}$. For a given m_1 and m_2, the conditional posterior distribution is

$$\pi(K_{m_1\to m_2} \,|\mathbf{B}, \boldsymbol{\theta}_g) \propto \pi(K_{m_1\to m_2})\times \prod_{\substack{j:d_j=m_1,\\ B_j=0}} \exp(-K_{m_1\to m_2}\, G_{m_1\to m_2}(T-t_j))\, K_{m_1\to m_2}^{|S_{j,m_2}|}, \tag{6.18}$$

which is sampled in a classic Metropolis-Hastings manner.

Sample $\mathbf{L}^{(k+1)}$

There are another M^2 entries in $\mathbf{L}$. For a given m_1 and m_2, the conditional posterior distribution is

$$\pi(L_{m_1\to m_2} \,|\mathbf{B}, \boldsymbol{\theta}_h) \propto \pi(L_{m_1\to m_2})\times \prod_{\substack{j:d_j=m_1,\\ B_j>0}} \exp(-L_{m_1\to m_2}\, H_{m_1\to m_2}(T-t_j))\, L_{m_1\to m_2}^{|S_{j,m_2}|}, \tag{6.19}$$

which is sampled according to Metropolis-Hastings.

Sample $\boldsymbol{\theta}_g^{(k+1)}$

The conditional posterior distribution for the parameters of $g_{m_1\to m_2}(\cdot)$ is

$$\pi(\theta_{g|m_1\to m_2} \,|\mathbf{B}, \mathbf{K}) \propto \pi(\theta_{g|m_1\to m_2})\times \prod_{\substack{j:d_j=m_1,\\ B_j=0}} \exp(-K_{m_1\to m_2}\, G_{m_1\to m_2}(T-t_j)) \prod_{i\in S_{j,m_2}} g_{m_1\to m_2}(t_i-t_j). \tag{6.20}$$

Again, these entries are independently sampled using Metropolis-Hastings.

Sample $\boldsymbol{\theta}_h^{(k+1)}$

Finally, the conditional posterior distribution for the parameters of $h_{m_1 \to m_2}(\cdot)$ is

$$\pi(\theta_{h|m_1\to m_2} \,|\mathbf{B}, \mathbf{L}) \propto \pi(\theta_{h|m_1\to m_2}) \times \prod_{\substack{j: d_j = m_1, \\ B_j > 0}} \exp(-L_{m_1\to m_2} H_{m_1\to m_2}(T - t_j)) \prod_{i \in S_{j,m_2}} h_{m_1\to m_2}(t_i - t_j). \tag{6.21}$$

These entries are independently sampled using Metropolis-Hastings.

6.4 Simulated Data Experiment

The goal of this section is to compare the Ancestor Hawkes to the classic Hawkes approach, highlighting the additional flexibility the Ancestor Hawkes model offers. To that end, we simulate a data set from the Ancestor Hawkes model and provide parameter estimates from both the Ancestor Hawkes and classic Hawkes model. We examine the posterior distributions obtained from each model. This permits a comparison of parameter estimates of similar parameters, such as the background rate. Moreover, we contrast the parameter estimates for the influence magnitudes, $\mathbf{K}$ in the classic Hawkes and $(\mathbf{K}, \mathbf{L})$ for the Ancestor Hawkes. To that end, we simulate data sets from each model using the respective posterior distributions and compare the models based on selected summary statistics (Meng, 1994). This emphasises the added flexibility of the Ancestor Hawkes model over to the classic approach.

6.4.1 Data

For this experiment, we use a simulated data set Y. This comes from an Ancestor Hawkes model in three dimensions, i.e. $M = 3$, and we set $T = 750$. We use a constant background rate and exponential influence kernel. Parameters are chosen as follows. $\mu_m = 0.05$ for $m = 1 \dots M$. $K_{m_1 m_2} = 0.6$ for $m_1, m_2 = 1 \dots M$. $L_{m_1 m_2} = 0.3$ where $m_1 = m_2$ and $L_{m_1 m_2} = 0.05$ where $m_1 \neq m_2$. $\beta_{m_1 m_2} = 2$ and $\gamma_{m_1 m_2} = 0.5$. The resulting data set contains 444 events. Figure 6.1 plots the events.

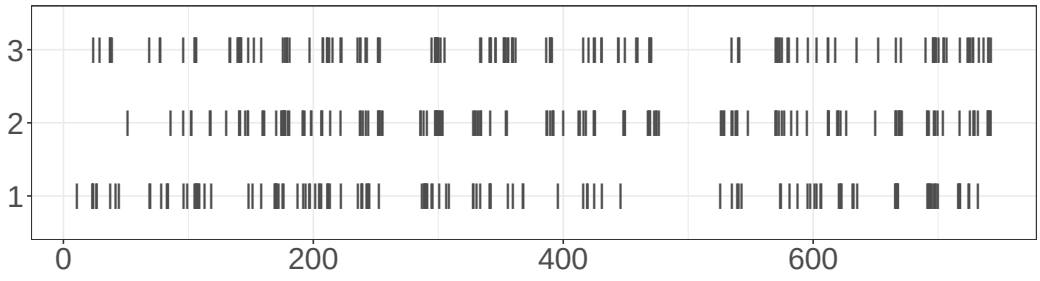


Figure 6.1: Events in the simulated data set for each dimension. Each vertical bar indicates an event.

6.4.2 Models

Both the classic Hawkes and Ancestor Hawkes models are used to provide posterior estimations based on the data set Y. For both models, we are utilising a constant background rate with parameters $\boldsymbol{\mu} = (\mu_1 \dots \mu_M)$. All influence kernels are exponential kernels, which are parameterised by $\boldsymbol{\theta}_g$ for the classic Hawkes and the immigrant-influence in the Ancestor Hawkes. Further, $\boldsymbol{\theta}_h$ parameterises the exponential kernel for the influence of triggered events in the Ancestor Hawkes.

Similar to Chapter 5 we limit the number of distinct parameters in the influence kernel. We assume that the shape is the same for all self-influences, and a different shape parameter is used for the cross-influences, i.e.

$$\begin{aligned} g_{j\to m}(x) &= \beta_{\text{diag}} \exp\left(-\beta_{\text{diag}}\, x\right), && \text{if } j = m, \\ g_{j\to m}(x) &= \beta_{\text{off}} \exp\left(-\beta_{\text{off}}\, x\right), && \text{if } j \neq m. \end{aligned} \tag{6.22}$$

Similarly, for the Ancestor Hawkes, define the following influence kernels.

$$\begin{aligned} h_{j\to m}(x) &= \gamma_{\text{diag}} \exp\left(-\gamma_{\text{diag}}\, x\right), && \text{if } j = m, \\ h_{j\to m}(x) &= \gamma_{\text{off}} \exp\left(-\gamma_{\text{off}}\, x\right), && \text{if } j \neq m. \end{aligned} \tag{6.23}$$

Again, this means that the same decay parameter is used for the self-influences, and similarly the cross-influences share one kernel parameter.

6.4.3 Prior Choice

We place the following priors on the parameters in the classic Hawkes setting:

$$\begin{aligned} \mu_m &\sim \text{Gamma}(2, 0.5), && \text{for } m = 1 \dots M, \\ K_{m_1\to m_2} &\sim \text{Uniform}(0, 0.9), && \text{for } m_1 = 1 \dots M, m_2 = 1 \dots M, \\ \beta_{\text{diag}} &\sim \text{Gamma}(1.5, 1), \\ \beta_{\text{off}} &\sim \text{Gamma}(1.5, 1). \end{aligned}$$

The Ancestor Hawkes employs the same priors as the classic Hawkes. Its additional priors are

$$\begin{aligned} L_{m_1\to m_2} &\sim \text{Uniform}(0, 0.9), && \text{for } m_1 = 1 \dots M, m_2 = 1 \dots M, \\ \gamma_{\text{diag}} &\sim \text{Gamma}(1.5, 1), \\ \gamma_{\text{off}} &\sim \text{Gamma}(1.5, 1). \end{aligned}$$

6.4.4 Results

This section contrasts the parameter estimates for both the Ancestor Hawkes and the classic Hawkes for data set Y. First, the posterior distributions of the background rate and the influence magnitudes are compared. Furthermore, we investigate the difference between the two models using summary statistics and posterior p-values (Meng, 1994). This highlights the added flexibility of the Ancestor Hawkes model as it captures aspects of the data that the classic Hawkes

model cannot describe as well.

We now consider posterior distributions from both models. We visually checked trace plots when fitting the models and were satisfied with the mixing behaviour for both of them. Figure 6.2 displays the posterior distributions for the background parameters for both models. It shows that the classic Hawkes estimates a higher background rate in all dimensions, whereas the Ancestor Hawkes posteriors are closer to the true value.

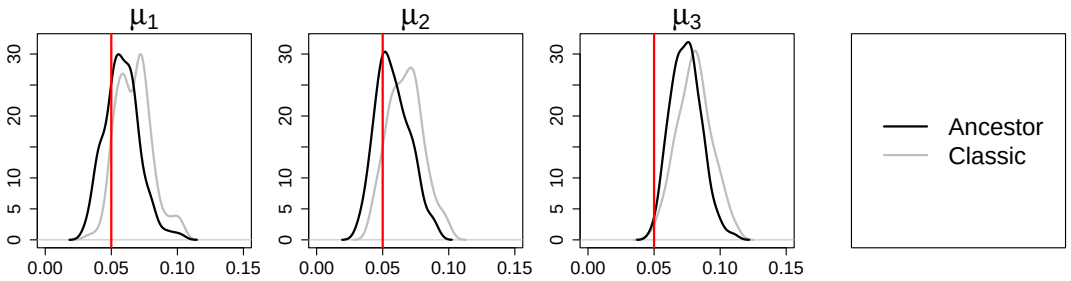


Figure 6.2: Posterior distribution estimates for the background rate for each dimension. The black line gives the posterior distribution in the Ancestor Hawkes model, grey line shows the posterior distribution for the classic Hawkes. The vertical red line indicates the true value.

Figure 6.3 gives the posterior distributions for all influence magnitude parameters. In the classic Hawkes, this is just the $\mathbf{K}$ parameter, for the Ancestor Hawkes both $\mathbf{K}$ and $\mathbf{L}$ are shown. Here it is evident that the posterior distributions for the classic Hawkes attempt to find a middle ground between the distributions of $\mathbf{K}$ and $\mathbf{L}$.

Summary Statistics

To compare the two models we make use of summary statistics that we originally developed for an ABC approach to Hawkes process modelling in Chapter 3, where the exact specifications of these summary statistics can be found. In addition, we also use the autocorrelation function (ACF) of the interevent times for different lags as an additional summary statistic for this comparison.

To that end, parameters are drawn from the posterior distributions and, based on these, new data sets are then sampled from the respective model. The distributions of summary statistics are compared to the true value of the summary statistics observed in Y using posterior p-values (Meng, 1994). This showcases the additional flexibility of the Ancestor Hawkes model beyond the classic Hawkes.

Figure 6.4 displays five summary statistics, which are calculated on all events across all dimensions. Three of those (top row) show that simulated data sets produce very similar distributions of summary statistics for both models. They also fit the true data well, as the posterior p-values are not extreme. The Ripley's K summary statistic (displayed bottom left) shows distinct distributions for the two models, with the Ancestor model providing a better fit to the data Y (posterior p-value in the Ancestor Hawkes is 0.27 compared to 0.08 in the classic Hawkes model). The result is similar for the autocorrelation of interevent times

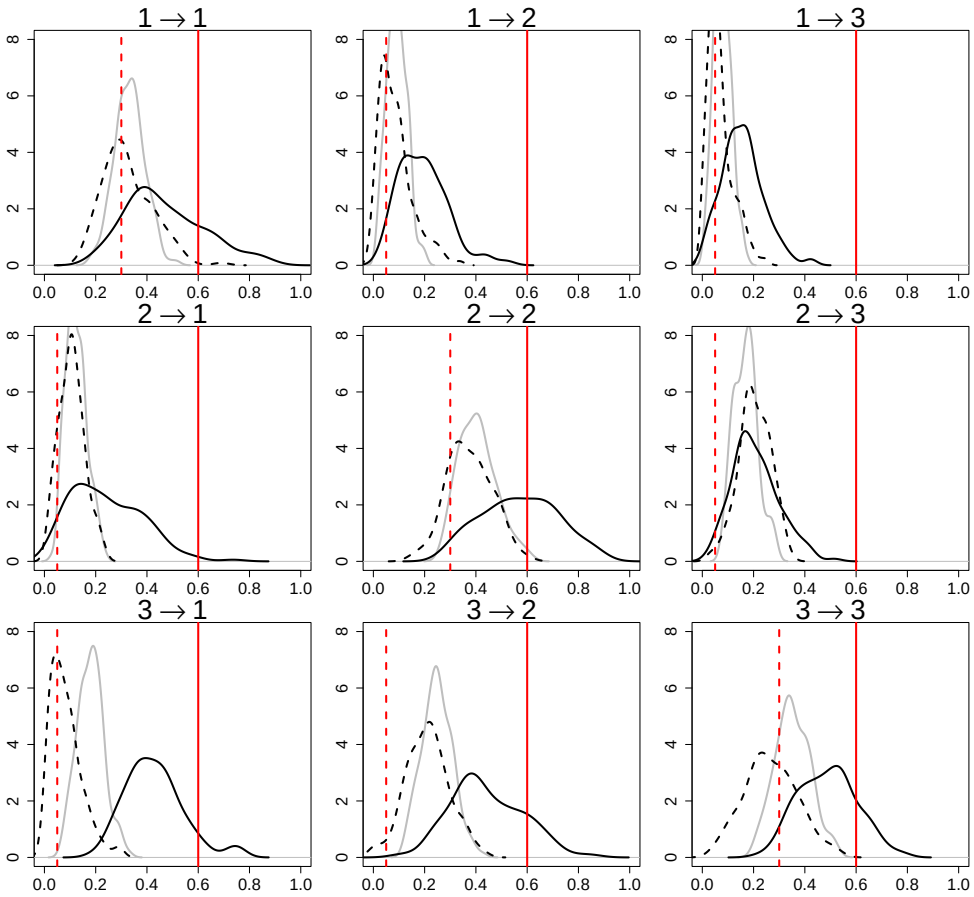


Figure 6.3: Posterior distribution estimates for the influence magnitude. The black solid line gives the posterior distribution for **K** in the Ancestor Hawkes model and the black dashed line represents posterior estimates for **L** in Ancestor Hawkes, grey line shows the posterior distribution of **K** in the classic Hawkes. It is evident that the classic Hawkes posteriors often attempt to find a middle ground between the posterior of **K** and **L** from the Ancestor model. The vertical red lines indicate the true values (solid for **K** and dashed for **L**) .

(bottom middle), with a posterior p-value of 0.05 in the Ancestor Hawkes model and < 0.01 for the classic approach.

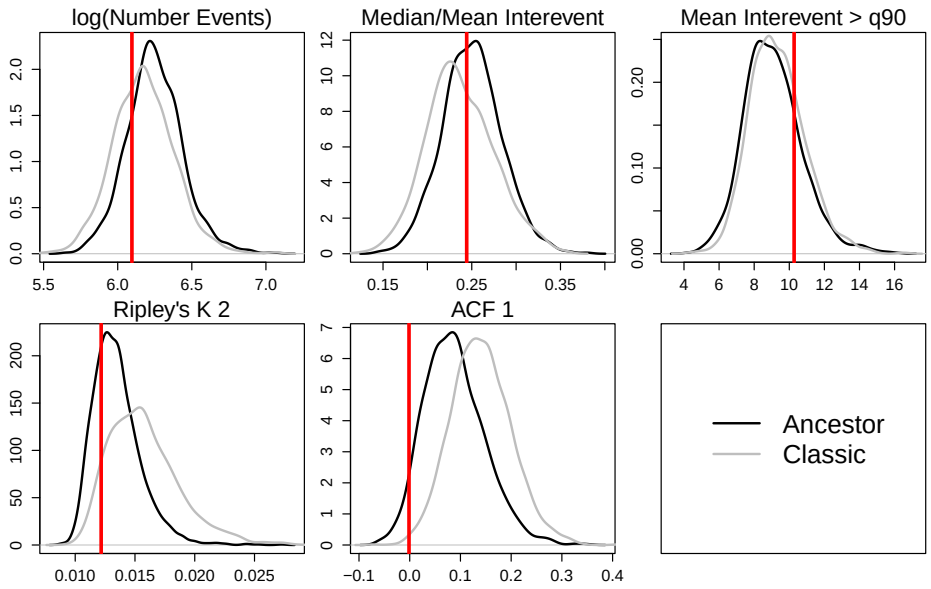


Figure 6.4: Comparison of distribution of summary statistics. The black line shows distribution when data is simulated from the Ancestor Hawkes model. The grey line gives the distribution for the classic Hawkes. The red vertical bar indicates the true value from the simulated data set Y. All summary statistics are calculated across the full data set. **Top**: Logarithm of number of events. Median of the interevent times divided by the mean of the interevent times. Mean of the interevent times that lie above their 90% quantile. **Bottom**: Ripley's K calculated with a window size of 2 (Ripley's K not to be confused with the parameter **K**). Autocorrelation function of the interevent times with lag 1.

This difference in the Ripley's K statistic between the two models is consistent across different data sets and a variety of window sizes. Similarly, the distinct distributions of the autocorrelation of interevent times persist for other choices of lags. The other summary statistics considered in Figure 6.4 describe "global" aspects of the data as they are averages over the whole data set. This suggests that both models are capable of capturing these overall trends in data that was generated from the Ancestor Hawkes model. However, data from the Ancestor Hawkes model displays additional complexities in the short-term distribution of events, which only the matching Ancestor Hawkes can capture. This indicates that the added flexibility of the Ancestor Hawkes is particularly pronounced in "local" patterns. The summary statistics in the top row all offer some kind of "global" average across the whole observational period (such as the logarithm of the number of events). In contrast, the Ripley's K and autocorrelation function examine what happens in much shorter intervals after each event (and averaging over those), which allows them to pick up different, more localised, patterns in the data.

6.5 Group Chat Data Application

This section gives an application of the Ancestor Hawkes model on data collected from a group chat setting. Both the data and the chosen model are described. The results emphasise both the flexibility and interpretability of our proposed approach.

6.5.1 Data

To highlight the benefits of the Ancestor Hawkes model we have collected data from a group chat setting. The data was downloaded from Facebook Messenger, where the chat was hosted, by one of the chat participants. All chat participants have given consent for their anonymised data to be analysed and the author has sought the required ethics approval through the University of Edinburgh.

The data comes from a chat group with 9 participants. Here, each message sent to the group chat is simultaneously received by all members and everyone has equal opportunity to participate. This setting is substantially different to the email data set examined by Miscouridou *et al.* (2018), where each email has exactly one receiver.

The full data set spans $10,705$ messages over approximately 3.5 years (from 11^{th} of October 2019 to 29^{th} of April 2023). The group chat was started with a subset of participants and additional members were added over time, but for most of the examined period, all 9 participants were part of the group chat. The data set only contains the sender and times at which a message was sent. No additional information (e.g. content of the message, whether it was a reply) is available. Figure 6.5 displays the total number of messages sent by each person. Figure 6.6 plots a subset of messages in a one-year period.

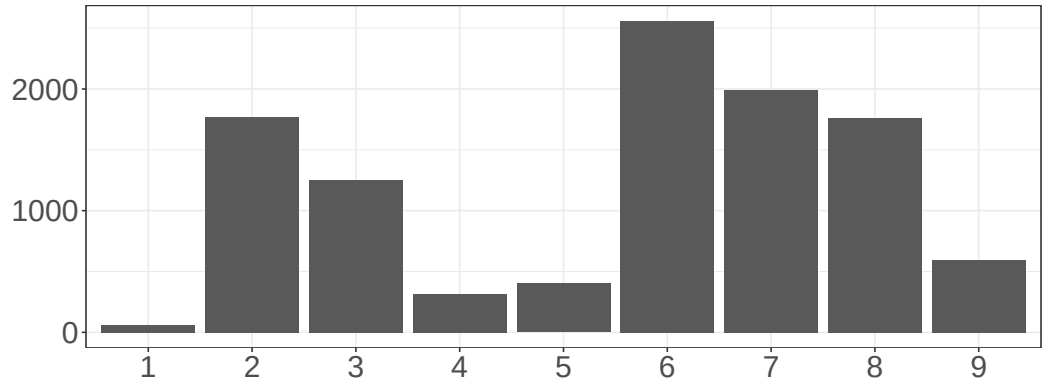


Figure 6.5: Total number of messages sent by each participant in the group chat.

First, we examine broad temporal patterns in the data. Figure 6.7 gives the number of messages sent per month. These remain somewhat stable across time, apart from a particularly high number of messages in March 2020. This coincides with the start of lockdowns across Europe (where all participants were based at that time) due to the Covid-19 pandemic.

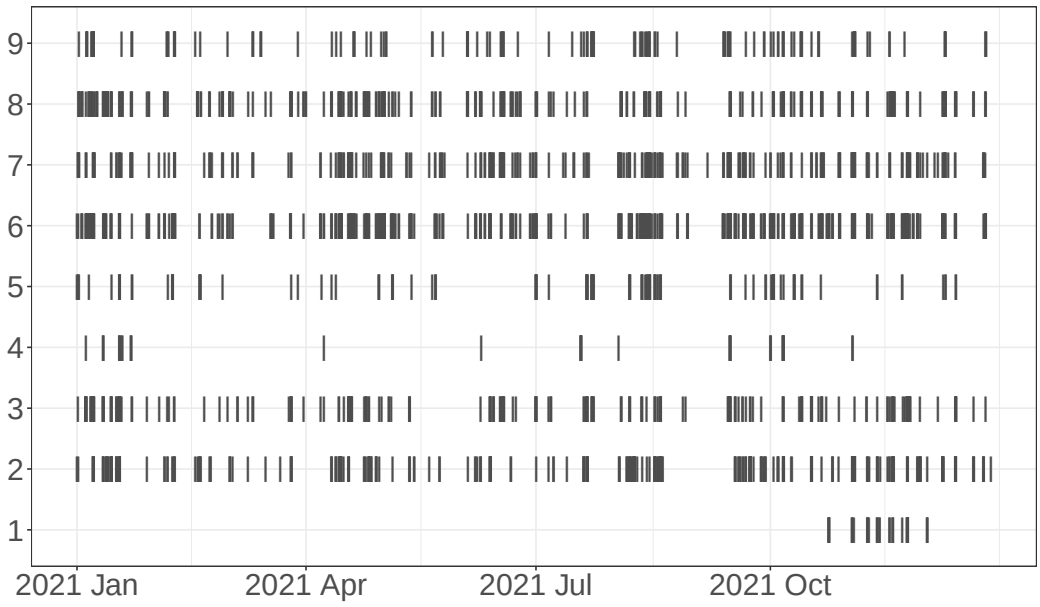


Figure 6.6: Messages sent by each participant in the group chat in 2021. All participants had already been added to the chat. Each vertical bar indicates a message sent.

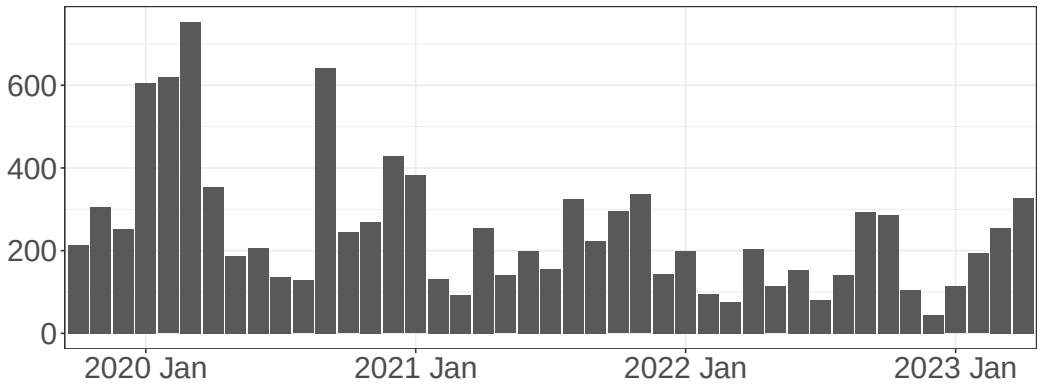


Figure 6.7: Number of messages sent per month.

Figure 6.8 shows the number of messages sent for each day of the week. The colour indicates which hour of the day a message was sent. The hours of the day are grouped into four categories each containing a quarter of the day. While there is a big difference within a day between working hours (biggest region in pink on the bottom) and late at night (barely visible orange on the top), these numbers stay similar across days of the week.

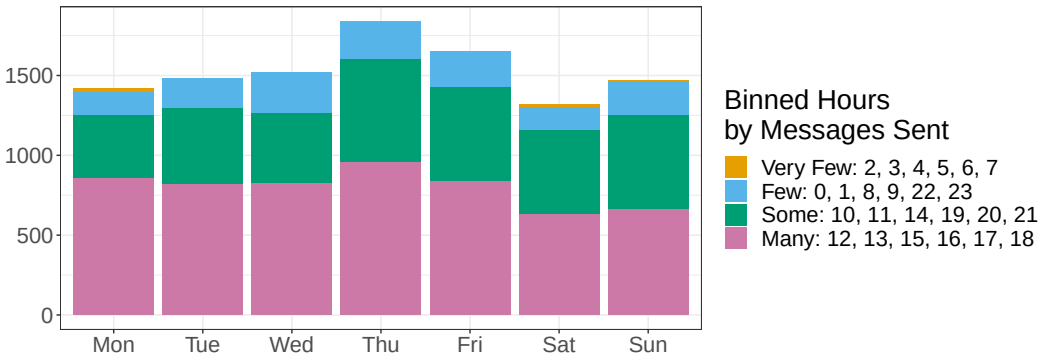


Figure 6.8: Number of messages sent per weekday. The colour indicates which hour of the day (grouped into four bins as described in the legend) the messages were sent. These bins are then used to construct the background rate as shown in Figure 6.9.

For our analysis we only model messages sent in 2021. All 9 participants were already part of the group chat, but we excluded two participants (Participant 1 and 4) from the analysis due to their low engagement. For them, there is little data to accurately estimate their influence parameters. Hence, their resulting posterior distributions would be heavily reliant on the prior, which would not produce interesting data-driven insights. Furthermore, their exclusion decreases the computational time. This leaves 7 participants with $2,680$ total messages in a 365 day window to be analysed.

6.5.2 Model

To model the temporal patterns of messages being sent we utilise an Ancestor Hawkes model where each participant is represented by a dimension and its respective intensity function. As a comparison model, we also fit a classic Hawkes process on the same data. Where not specified otherwise, the same settings are applied.

For both models, a non-constant background rate is used. Hence, the intensity function of the Ancestor Hawkes in dimension m is

$$\lambda_m(t \mid \boldsymbol{\theta}, \mathbf{B}) = \mu_m(t) + \sum_{\substack{i:t_i<t,\\ B_i=0}} K_{d_i \to m}\, g_{d_i \to m}(t - t_i) + \sum_{\substack{j:t_j<t,\\ B_j>0}} L_{d_j \to m}\, h_{d_j \to m}(t - t_j). \tag{6.24}$$

We model the message-sending pattern from seven participants. Note that the subsequent plots retain the numbering of participants from 1 to 9 as displayed in Figure 6.6, but with Participants 1 and 4 excluded. Hence the estimation problem is in 7 dimensions and we use $\mathcal{M} = \{2, 3, 5, 6, 7, 8, 9\}$ to select the relevant indices. We now specify the choices made for the background rate and the influence kernel.

Background Rate

We choose a background rate of the form

$$\mu_m(t) = c_m \, b(t), \tag{6.25}$$

for $m \in \mathcal{M}$ where $b(t) > 0$ accounts for trends within a day and $c_m > 0$ is a participant-specific scalar. Critically, $b(t)$ is produced outside of the model and only the parameter c_m is estimated jointly with the other parameters. This is paralleling the approach from Chapter 5 and Helmstetter *et al.* (2006), where plug-in estimates are used for parts of the background rate.

Here, $b(t)$ captures the fluctuation of events within a day due to the time, where more events are sent during the day than in the middle of the night. To that extent, we group the hours of the day into quartiles according to the number of messages sent in that hour over the complete 3.5 years. Therefore there are four different levels of background rate occurring each day. Figure 6.9 shows both their grouping and the respective level of $b(t)$.

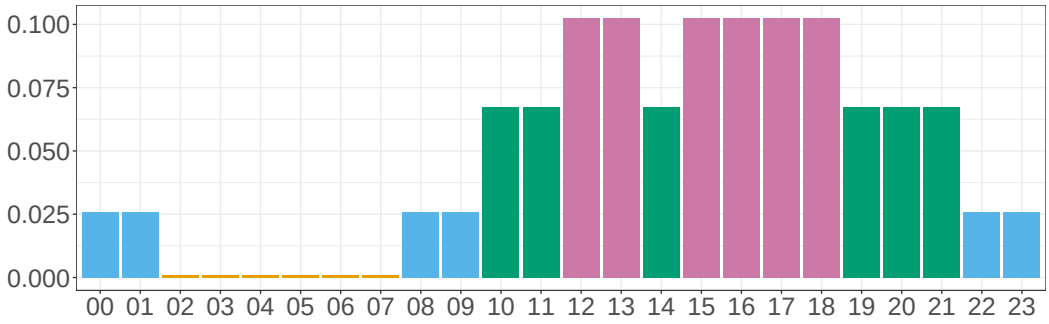


Figure 6.9: Values of $b(t)$, the part of the background rate, which is estimated outside of the model. The colours indicate which hours are grouped together, i.e. there are only four different values the background rate can take. This parallels the bins shown in Figure 6.8.

Influence Kernel

Similar to the simulated data example and Chapter 5 we limit the number of distinct parameters in the influence kernel. We assume that the shape is the same for all self-influences, and a different shape parameter is used for the cross-

influences, such that

$$\begin{aligned} g_{j\to m}(x) &= \beta_{\text{diag}} \exp\left(-\beta_{\text{diag}}\, x\right), && \text{if } j = m, \\ g_{j\to m}(x) &= \beta_{\text{off}} \exp\left(-\beta_{\text{off}}\, x\right), && \text{if } j \neq m. \end{aligned} \tag{6.26}$$

A similar assumption is made for the Ancestor Hawkes:

$$\begin{aligned} h_{j\to m}(x) &= \gamma_{\text{diag}} \exp\left(-\gamma_{\text{diag}}\, x\right), && \text{if } j = m, \\ h_{j\to m}(x) &= \gamma_{\text{off}} \exp\left(-\gamma_{\text{off}}\, x\right), && \text{if } j \neq m. \end{aligned} \tag{6.27}$$

6.5.3 Prior Choice

We place the following priors on the parameters in the classic Hawkes setting:

$$\begin{aligned} c_m &\sim \text{Gamma}(2, 0.5), && \text{for } m \in \mathcal{M}, \\ K_{m_1\to m_2} &\sim \text{Uniform}(0, 0.9), && \text{for } m_1 \in \mathcal{M}, m_2 \in \mathcal{M}, \\ \beta_{\text{diag}} &\sim \text{Gamma}(1.5, 1), \\ \beta_{\text{off}} &\sim \text{Gamma}(1.5, 1). \end{aligned}$$

The Ancestor Hawkes employs the same priors as the classic Hawkes. Its additional priors are

$$\begin{aligned} L_{m_1\to m_2} &\sim \text{Uniform}(0, 0.9), && \text{for } m_1 \in \mathcal{M}, m_2 \in \mathcal{M}, \\ \gamma_{\text{diag}} &\sim \text{Gamma}(1.5, 1), \\ \gamma_{\text{off}} &\sim \text{Gamma}(1.5, 1). \end{aligned}$$

6.5.4 Results

We now present the results of the Ancestor Hawkes model and compare them to the classic Hawkes. Due to the relatively large number of parameters we only present selected plots here. Additional figures can be found in Appendix B. We then showcase the goodness of fit for both models based on summary statistics.

We obtained 3000 raw samples of the posterior, of which 50% were discarded as burn-in. The samples were then thinned such that only 150 samples were kept. This took approximately 5 hours on a standard laptop. As above, trace plots were visually examined when fitting these models and we were satisfied with their mixing behaviour.

First, we investigate the posterior estimates for the influence magnitude in more detail. The top row of Figure 6.10 compares the posterior estimates for $K_{7\to m}$ and $L_{7\to m}$, where $m = 7$ (self-influence) is displayed in a dotted line. This describes the change in the other participants' intensity functions due to an immigrant message ($K_{7\to m}$) or a follow-up message ($L_{7\to m}$) from the 7^{th} participant. It is evident that immigrant messages solicit more responses. This is true across all participants and in line with our understanding of messaging dynamics.

The bottom row of Figure 6.10 depicts the converse. Here, $K_{m\to 7}$ and $L_{m\to 7}$ show the response of the 7^{th} participant to an immigrant and triggered message

from someone else. It highlights the fact that this participant is more likely to respond to an immigrant message from Participant 3 than participant 5 ($\hat{K}_{3\to7} = 0.44$ versus $\hat{K}_{5\to7} = 0.13$). In addition, they are also likely to reply to their own message with $\hat{K}_{7\to7} = 0.53$ and $\hat{L}_{7\to7} = 0.27$. Hence, each immigrant message they send is, on average, followed by 0.53 of their own triggered messages. Such a comparably larger self-influence can be observed for most participants. This parallels the idea that people might send multiple messages about the same topic in rapid succession.

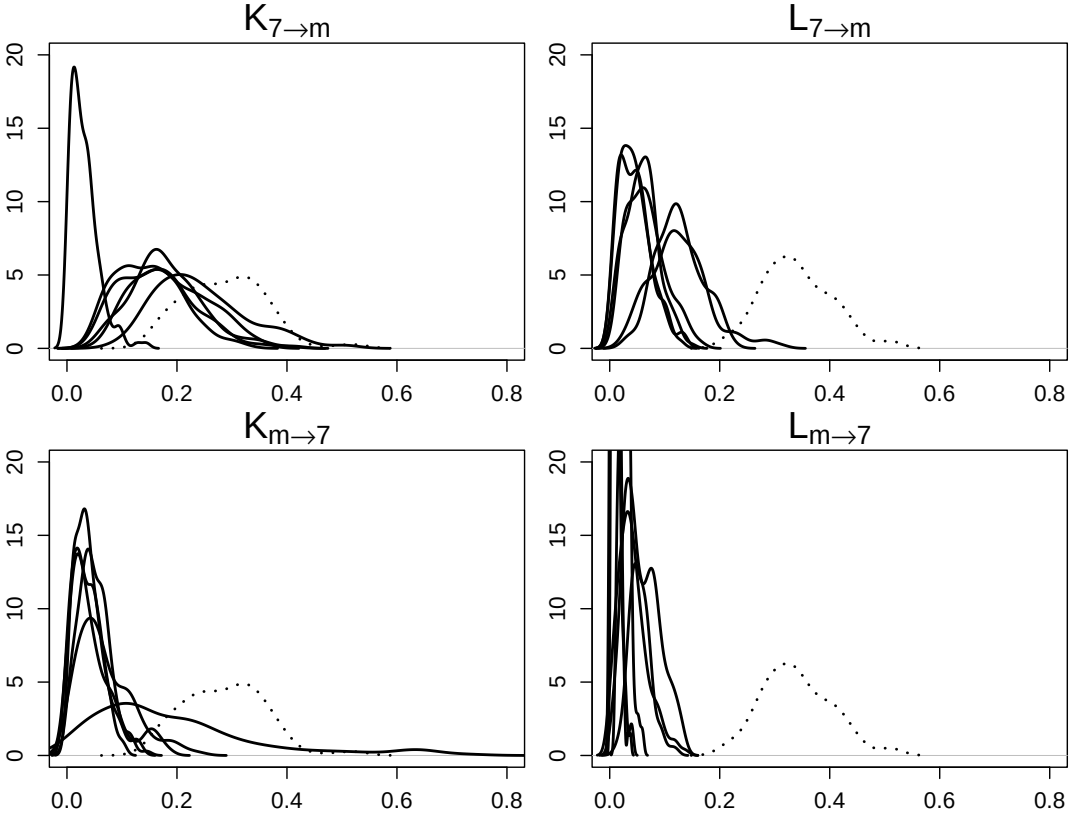


Figure 6.10: Posterior distribution estimates for the influence magnitude parameters in the Ancestor Hawkes model for the 7^{th} participant. The dotted line indicates that $m = 7$, i.e. self-influence. This plot serves as an overview, each posterior distribution is shown separately in Appendix B. To plot all figures on the same scale it was unavoidable that two distributions of the bottom-right plot were cut off as they are rather concentrated around zero.

Comparison to Classic Hawkes

This section compares the aforementioned results obtained via the Ancestor Hawkes model to that of the classic Hawkes. Figure 6.11 shows the posterior distribution of the background rate scalar c_m of each participant in the Ancestor Hawkes and the classic Hawkes model. Across all participants, the Ancestor Hawkes model estimates a lower value for this scalar.

Interestingly, Participant 2 and 6 have rather similar values ($\hat{c}_2 = 0.20$ and $\hat{c}_6 = 0.22$ in Ancestor Hawkes) but Participant 6 has a total of 735 messages in

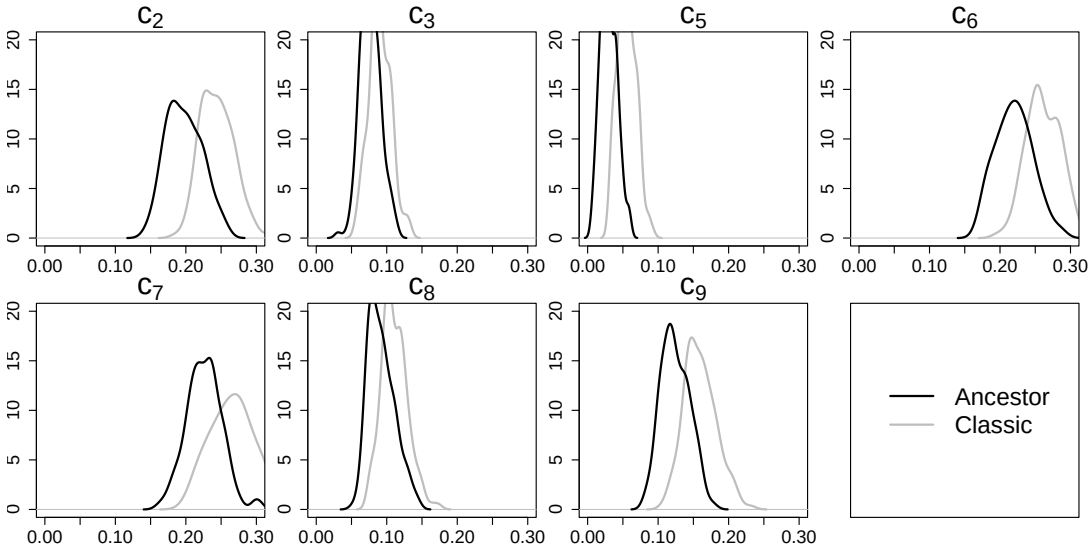


Figure 6.11: Posterior distribution estimates for the scalar of the background rate for each participant. Participant 1 and 4 were excluded due to their low engagement. The black line gives the posterior distribution in the Ancestor Hawkes model, grey line shows the posterior distribution for the classic Hawkes.

the analysed data set, whereas Participant 2 has only sent 407 messages. This indicates that Participant 2 is more likely to send an immigrant message that starts a conversation cascade, whereas Participant 6 is more likely to respond to messages.

Figure 6.12 compares the posterior distributions for $\mathbf{K}$ and $\mathbf{L}$ from the Ancestor Hakes to the posterior of $\mathbf{K}$ from the classic Hawkes. For this plot, we chose to investigate Participant 2. Posterior distributions for all participants are available in Appendix B.1. Since the classic Hawkes cannot assign a different influence to immigrant and triggered events, it often sits in between the two curves from the Ancestor Hawkes model, such as $K_{2\to6}$. In other cases, such as $K_{2\to2}$, it mimics the lower posterior of $\mathbf{L}$, which is compensated for by the higher background rate estimates.

Summary Statistics

As in Section 6.4, we use summary statistics from simulated data by the respective models to compare the two approaches to the true value of the summary statistics observed in the group chat data set.

Figure 6.13 compares the distribution of five selected summary statistics from both models to the true value. More summary statistics, as well as distributions for each dimension separately, are available in Appendix B.2. It is evident that there are some aspects of the data that both the Ancestor Hawkes and the classic Hawkes capture well. For the summary statistics plotted on the top left and top right, both models give distributions where the observed value does not land in the extremes. This indicates a reasonable fit regarding these aspects of the data.

For the summary statistic shown in the top middle, this is not the case. Here, both models perform poorly as the observed value is smaller than all values

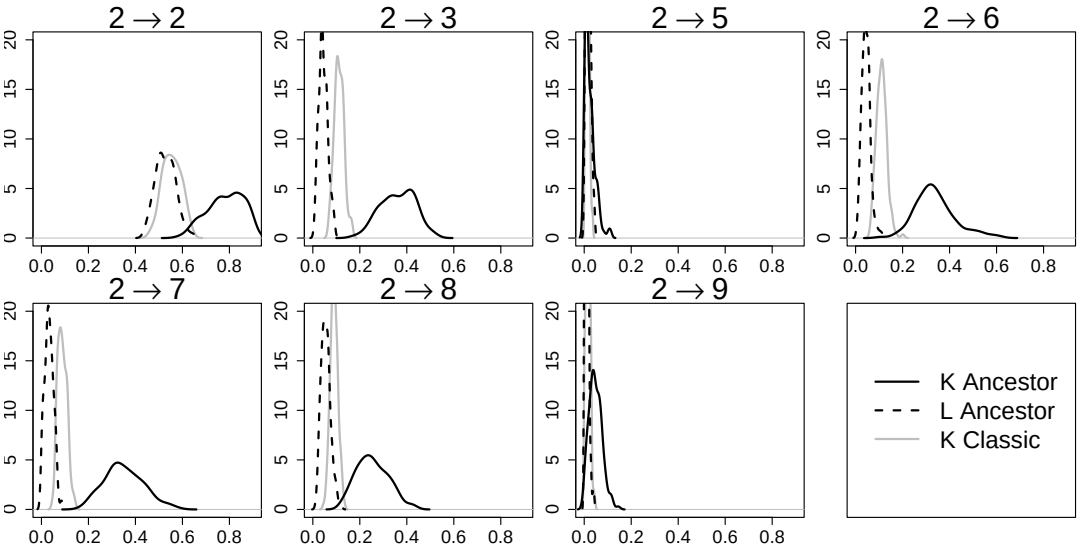


Figure 6.12: Posterior distribution estimates for the influence magnitude in response to a message of Participant 2. Participant 1 and 4 were excluded due to their low engagement. The black solid line gives the posterior distribution for **K** in the Ancestor Hawkes model and the black dashed line represents posterior estimates for **L** in Ancestor Hawkes, grey line shows the posterior distribution of **K** in the classic Hawkes.

obtained from simulated data, i.e. an extreme posterior p-value. While both show a weak match to the true data, the distribution curve from the classic Hawkes is at least closer to the true value. However, this summary statistic is particularly susceptible to large interevent times, which increases the mean of interevent times. As evident in Figure 6.6, there are longer periods where no messages are sent at all, which both the Ancestor and the classic Hawkes model fail to depict. To alleviate this, a more elaborate background rate could be used, which takes, for example, bank holidays or summer/winter breaks into account. This would allow both models to incorporate prolonged periods of no messages and could improve the posterior p-values for this summary statistic.

Finally, the summary statistics at the bottom show a difference between the two models. The plot on the bottom left depicts the distribution of the Ripley's K statistic for a window size of 2 hours. This broadly measures how many messages are sent in the 2 hours after each message. Here, the Ancestor Hawkes model provides a rather good match to the data (posterior p-value of 0.56), whereas the observed value is unlikely in the classic Hawkes model (posterior p-value is 0.01). Similarly, the autocorrelation of interevent times with lag 1 observed is more extreme than most simulated values in the classic Hawkes model (posterior p-value of 0.01) but is well matched by the Ancestor Hawkes model (posterior p-value of 0.33). This indicates that there are some aspects of the data that only the Ancestor Hawkes model is able to capture. This aligns well with our experiment in Section 6.4 that highlights the fact that the flexibility of the Ancestor Hawkes is most pronounced in the Ripley's K and the ACF statistics.

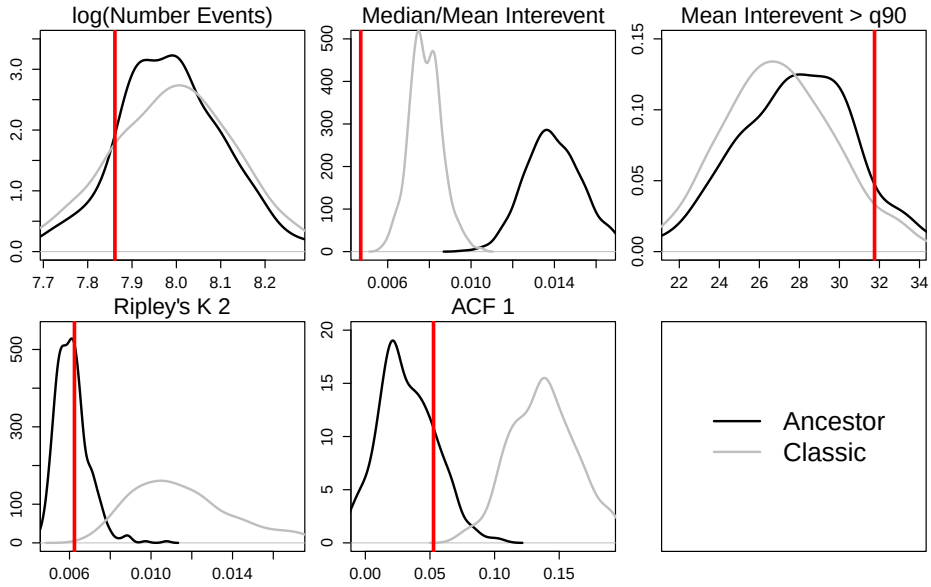


Figure 6.13: Comparison of distribution of summary statistics. The black line shows distribution when data is simulated from the Ancestor Hawkes model. The grey line gives the distribution for the classic Hawkes. The red vertical bar indicates the true value from the observed group chat data. All summary statistics are calculated across the full data set. **Top**: Logarithm of number of events. Median of the interevent times divided by the mean of the interevent times. Mean of the interevent times that lie above their 90% quantile. **Bottom**: Ripley's K calculated with a window size of 2 (Ripley's K not to be confused with the parameter **K**). Autocorrelation function of the interevent times with lag 1.

6.6 Discussion

In this chapter, we have introduced a novel Hawkes process model. The Ancestor Hawkes permits a different influence for immigrant and triggered events, which adds flexibility and mirrors real-world mechanics, such as in the considered group chat setting. The No-Cascade Hawkes additionally limits the influence of triggered events into other dimensions, which prohibits cascades through many dimensions. With that, the proposed model eases the otherwise difficult Granger-causal interpretation, where it becomes difficult to differentiate a direct influence ('$j \to m$') and an indirect one ('$j \to k \to m$') (Eichler, 2013). The Ancestor Hawkes makes this more explicit as it distinguishes direct and indirect influences.

We highlight the interpretability and flexibility of the Ancestor Hawkes in a simulation study that emphasises the fact that the Ancestor Hawkes can express particular local patterns (as measured by Ripley's K and the ACF) that the classic Hawkes cannot describe as well. We then showcase the Ancestor Hawkes model on group chat data and compare it to a classic Hawkes approach. While both examined models fail to capture all aspects of the group chat data, the Ancestor Hawkes model provides a better fit for summary statistics based on Ripley's K and the ACF, which represents the clustering of events in a short time frame. Further analysis should include an improved background rate.

The data set analysed in this chapter was collected specifically for this line of research. Crucially, it only contains the times at which messages were sent and who the sender was. No other information is recorded. For future work, additional covariates could be incorporated into the existing model. For example, most messaging platforms now allow users to "reply" to a specific message. When this function is used it can directly inform the branching structure. Different influences depending on the type of message, such as text, GIF, emoji only, or picture, could also be considered by the model. In addition, "emoji reactions" (replies of a single emoji displayed in the corner of the original message) could be included – either as messages in their own right or as marks similar to the ETAS setting. Finally, the content of each message is a central source of additional information, which opens the door to semantic analysis.

This work represents the first introduction of the Ancestor Hawkes and No-Cascade Hawkes model. It would be interesting to see these models being applied in a variety of other settings, such as trading on the stock market or neural activity. Similarly, this is one of the first instances of group chat data modelling and we would like to compare our proposed approach to other techniques successfully used in adjacent disciplines, such as in-person conversations between different people (Tan *et al.*, 2018).

Chapter 7

Conclusion

This thesis investigates various aspects of Bayesian modelling for Hawkes process, spanning both estimation and application. On one hand, we enable and improve estimation, for example when data is missing or when inhibition is incorporated into the model. On the other hand, we show novel applications of Hawkes processes to product cannibalisation and group chat data. We now provide a summary of our work and highlight avenues for future research.

After reviewing the background on Hawkes processes in Chapter 2, Chapter 3 examined Hawkes process estimation under missing data. Most commonly used methods assume complete and correctly observed data, which renders them inapplicable or biased when events are missing from the record. In addition, the likelihood of the model with missing data becomes intractable, such that MCMC methods, that rely on the evaluation of the likelihood, cannot be used. Our contribution is a tailor-made ABC-based method that provides an estimation of the posterior distribution.

Chapter 4 took a closer look at Hawkes processes with both excitation and inhibition. Here, we reviewed methods to ensure a non-negative intensity and suggested exact and approximate ways to integrate the intensity. In addition, we introduced a new criterion for stability when inhibition is present, as existing conditions were unnecessarily strict.

Based on these findings, Chapter 5 employed a multivariate Hawkes process with inhibition to model product cannibalisation. This allowed us to uncover product cannibalisation in two real-data examples for two product classes using the data provided by our industry partner. For this novel application of Hawkes processes, we proposed a prior that is independent of the dimensionality of the problem based on a reparametrisation of the model. We also showed that the Hawkes process with inhibition performed better out-of-sample than competing approaches without inhibition.

Chapter 6 introduced two variants of the classic Hawkes process model that use the latent branching structure in their parameter structure. We called these Ancestor and No-Cascade Hawkes. We presented an efficient estimation procedure and showcased their flexibility and interpretability on simulated and real-life data. For the latter, we have collected a data set from a group chat setting, where the Ancestor Hawkes approach permitted a more nuanced analysis of the messaging patterns between participants.

Future work could neatly sit in the intersections of the different topics examined in this thesis. Chapter 3 considers an ABC approach for missing data in a univariate, excitation-only scenario. This could be extended to cover a multivariate approach, where the "closeness" of observed and simulated data sets has to be assessed not only overall, but potentially for each dimension separately. This warrants an adapted distance function to ensure the acceptance rate does not become too small.

In addition, ABC for Hawkes processes with inhibition could be explored, similar to repulsive spatial point processes by Shirota and Gelfand (2017). Our introduced summary statistics might have to be adjusted or extended to account for inhibiting influences. Using ABC on complete data when inhibition is present would also circumvent the costly evaluation of the likelihood (similar to Ertekin *et al.*, 2015). Furthermore, it would be interesting to examine whether ABC approaches could be used to speed up computation for any multivariate Hawkes processes as they avoid cumbersome likelihood calculations.

We have demonstrated in Chapter 3 that our proposed summary statistics do a good job of capturing many aspects of data coming from a Hawkes process. Hence, these summary statistics may also be useful when assessing goodness of fit for Hawkes process models. We have utilised these summary statistics for that purpose already in Chapter 6.

There are also parallels between missing data and inhibition, as inhibition events "prevent" other events from happening/being recorded. Hence, Hawkes processes with inhibition can be viewed as excitation-only Hawkes processes with missing data. The attractive advantage of this interpretation is the preserved branching structure, which might be desirable for computational or interpretational reasons. Our work on both missing data and inhibition could serve as a starting point for this area of research.

Chapter 4 examines stability for Hawkes processes with inhibition. Additional work into stability could encompass stability for the Ancestor and No-Cascade Hawkes models we propose in Chapter 6.

Investigating whether common data sets where classic Hawkes processes are successfully used (such as earthquakes or crime) would benefit from the Ancestor Hawkes approach is another area of future work. Furthermore, it would be of interest to develop an Ancestor-type model under inhibition. This presents a challenge, as the classic branching structure formulation is not available when inhibition is present. Nevertheless, a model that can account for different types of influences could be beneficial, for example for the product cannibalisation application.

Finally, we have showcased the usefulness of a variety of Hawkes process models for two applications, namely product cannibalisation and group chat data. Given the versatility and broad applicability of Hawkes process models, we would like to explore more use cases such as neural inhibition in biology, stock tradings in economics, and various product sale scenarios.

Bibliography

Aguilar-Palacios, C., Muñoz-Romero, S. and Rojo-Álvarez, J. L. (2021). Causal quantification of cannibalization during promotional sales in grocery retail. *IEEE Access*, **9**, 34078–34089.

Aryal, N. R. and Jones, O. D. (2020). Fitting the Bartlett–Lewis rainfall model using approximate Bayesian computation. *Mathematics and Computers in Simulation*, **175**, 153–163.

Atasu, A., Guide, V. D. R. and Van Wassenhove, L. N. (2010). So what if remanufacturing cannibalizes my new product sales? *California Management Review*, **52**(2), 56–76.

Bacry, E. and Muzy, J.-F. (2016). First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, **62**(4), 2184–2202.

Bacry, E., Mastromatteo, I. and Muzy, J.-F. (2015). Hawkes Processes in finance. *Market Microstructure and Liquidity*, **1**(01), 1550005.

Bacry, E., Jaisson, T. and Muzy, J. (2016). Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, **16**(8), 1179–1201.

Bacry, E., Bompaire, M., Gaïffas, S. and Muzy, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, **21**(50), 1–32.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

Bekal, G. and Bari, M. (2021). An XGBoost-based forecasting framework for product cannibalization. *arXiv*.

Bernton, E., Jacob, P. E., Gerber, M. and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B*, **81**(2), 235–269.

Bessy-Roland, Y., Boumezoued, A. and Hillairet, C. (2021). Multivariate Hawkes process for cyber insurance. *Annals of Actuarial Science*, **15**(1), 14–39.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*.

Blundell, C., Beck, J. and Heller, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates Inc.

Bonnet, A., Herrera, M. M. and Sangnier, M. (2021). Maximum likelihood estimation for Hawkes processes with self-excitation or inhibition. *Statistics & Probability Letters*, **179**, 109214.

Bonnet, A., Martinez Herrera, M. and Sangnier, M. (2023). Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, **33**(4), 91.

Bremaud, P. and Massoulie, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, **24**(3), 1563–1588.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. IMS.

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V. and Rousseau, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of Covid-19. *PLOS ONE*, **16**(4), e0250015.

Cavaliere, G., Lu, Y., Rahbek, A. and Stærk-Østergaard, J. (2023). Bootstrap inference for Hawkes and general point processes. *Journal of Econometrics*, **235**(1), 133–165.

Chen, F. and Hall, P. (2016). Nonparametric Estimation for self-exciting point processes – a parsimonious approach. *Journal of Computational and Graphical Statistics*, **25**(1), 209–224.

Child, P., Diederichs, R., Sanders, F.-H. and Wisniowski, S. (1991). SMR forum: the management of complexity. *Sloan Management Review*, **33**(1), 73–80.

Copulsky, W. (1976). Cannibalism in the marketplace. *Journal of Marketing*, **40**(4), 103–105.

Costa, M., Graham, C., Marsalle, L. and Tran, V.-C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, **52**(3), 879–915.

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman and Hall/CRC.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2nd edition.

De Giovanni, P. and Ramani, V. (2018). Product cannibalization and the effect of a service strategy. *Journal of the Operational Research Society*, **69**(3), 340–357.

Del Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, **22**(5), 1009–1020.

Desai, P. S. (2001). Quality segmentation in spatial markets: when does cannibalization affect product line design? *Marketing Science*, **20**(3), 265–283.

Deutsch, I. and Ross, G. J. (2020). ABC learning of Hawkes processes with missing or noisy event times. *arXiv*.

Deutsch, I. and Ross, G. J. (2022). Bayesian estimation of multivariate Hawkes processes with inhibition and sparsity. *arXiv*.

Eichler, M. (2013). Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**(1997), 20110613.

Eichler, M., Dahlhaus, R. and Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, **38**(2), 225–242.

Embrechts, P., Liniger, T. and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, **48**(A), 367–378.

Ertekin, Ş., Rudin, C. and McCormick, T. H. (2015). Reactive point processes: a new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, **9**(1), 122–144.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B*, **74**(3), 419–474.

Fukumizu, K., Song, L. and Gretton, A. (2013). Kernel Bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, **14**(1), 3753–3783.

Garnier, R. (2022). Concurrent neural network: a model of competition between times series. *Annals of Operations Research*, **313**(2), 945–964.

Gelfand, I. (1941). Normierte Ringe. *Rech. Math. [Mat. Sbornik]*, **9**(51), 3–24.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Ghose, A., Smith, M. D. and Telang, R. (2006). Internet exchanges for used books: an empirical analysis of product cannibalization and welfare impact. *Information Systems Research*, **17**(1), 3–19.

Guidolin, M. and Guseo, R. (2020). Has the iPhone cannibalized the iPad? An asymmetric competition model. *Applied Stochastic Models in Business and Industry*, **36**(3), 465–476.

Guo, G., Wang, C., Chen, J., Ge, P. and Chen, W. (2019). Who is answering whom? Finding "reply-to" relations in group chats with deep bidirectional LSTM networks. *Cluster Computing*, **22**, 2089–2100.

Gutmann, M. U., Dutta, R., Kaski, S. and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, **28**(2), 411–425.

Guttorp, P., Illian, J., Kostensalo, J., Kuronen, M., Myllymäki, M., Särkkä, A. and Thorarinsdottir, T. L. (2023). What you see is not what is there: mechanisms, models, and methods for point pattern deviations. *arXiv*.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**(1), 83–90.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, **11**(3), 493–503.

Helmstetter, A. and Sornette, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research: Solid Earth*, **107**(B10), ESE 10–1–ESE 10–21.

Helmstetter, A., Kagan, Y. Y. and Jackson, D. D. (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *Bulletin of the Seismological Society of America*, **96**(1), 90–106.

Jovanović, S., Hertz, J. and Rotter, S. (2015). Cumulants of Hawkes point processes. *Physical Review E*, **91**(4), 042802.

Kamakura, W. A. and Srivastava, R. K. (1984). Predicting choice shares under conditions of brand interdependence. *Journal of Marketing Research*, **21**(4), 420–434.

Kolev, A. A. and Ross, G. J. (2019). Inference for ETAS models with non-Poissonian mainshock arrival times. *Statistics and Computing*, **29**(5), 915–931.

Kong, E. L. (2015). *Cannibalization Effects of Products in Zara's Stores and Demand Forecasting*. Ph.D. thesis, Massachusetts Institute of Technology.

Kong, Q., Calderon, P., Ram, R., Boichak, O. and Rizoiu, M.-A. (2023). Interval-censored transformer Hawkes: detecting information operations using the reaction of social systems. In *Proceedings of the ACM Web Conference 2023*, pages 1813–1821. Association for Computing Machinery.

Laub, P. J., Taimre, T. and Pollett, P. K. (2015). Hawkes processes. *arXiv*.

Laub, P. J., Lee, Y. and Taimre, T. (2021). *The Elements of Hawkes Processes*. Springer.

Le, T. (2018). A multivariate Hawkes process with gaps in observations. *IEEE Transactions on Information Theory*, **63**(3), 1800–1811.

Lemonnier, R. and Vayatis, N. (2014). Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases – Part II*, pages 161—-176. Springer.

Li, H., Li, H. and Bhowmick, S. S. (2020). BRUNCH: branching structure inference of hybrid multivariate Hawkes processes with application to social media. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference*, pages 553–566. Springer.

Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y. and Song, L. (2018). Learning temporal point processes via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates Inc.

Li, W. and Fearnhead, P. (2017). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika*, **105**, 301–318.

Linderman, S. W., Wang, Y. and Blei, D. M. (2017). Bayesian inference for latent Hawkes processes. In *Advances in Approximate Bayesian Inference Workshop, 31st Conference on Neural Information Processing Systems*. Curran Associates Inc.

Lindsten, F., Schön, T. and Jordan, M. (2012). Ancestor sampling for particle Gibbs. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Liniger, T. J. (2009). *Multivariate Hawkes Processes*. Ph.D. thesis, ETH Zurich.

Lu, X. and Abergel, F. (2018). High-dimensional Hawkes processes for limit order books: modelling, empirical analysis and numerical calibration. *Quantitative Finance*, **18**(2), 249–264.

Mannell, K. (2020). Plural and porous: reconceptualizing the boundaries of mobile messaging group chats. *Journal of Computer-Mediated Communication*, **25**(4), 274–290.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, **22**(6), 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100**(26), 15324–15328.

Markwick, D. (2020). *Bayesian Nonparametric Hawkes Processes with Applications*. Ph.D. thesis, University College London.

McCullough, K. and Ebrahimi, N. (2018). Approximate Bayesian computation for censored data and its application to reliability assessment. *IISE Transactions*, **50**(5), 419–430.

Mei, H. and Eisner, J. (2017). The neural Hawkes process: a neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates Inc.

Mei, H., Qin, G. and Eisner, J. (2019). Imputing missing events in continuous-time event streams. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4475–4485. PMLR.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, **22**(3), 1142–1160.

Miscouridou, X., Caron, F. and Teh, Y. W. (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates Inc.

Mohler, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics*, **7**(3), 1525–1539.

Molkenthin, C., Donner, C., Reich, S., Zöller, G., Hainzl, S., Holschneider, M. and Opper, M. (2022). GP-ETAS: semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model. *Statistics and Computing*, **32**(2), 29.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in Applied Probability*, **37**(3), 629–646.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, **27**(1), 23–31.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, **83**(401), 9–27.

Okorie, O., Obi, M., Russell, J., Charnley, F. and Salonitis, K. (2021). A triple bottom line examination of product cannibalisation and remanufacturing: a review and research agenda. *Sustainable Production and Consumption*, **27**, 958–974.

Omi, T., Ogata, Y., Hirata, Y. and Aihara, K. (2014). Estimating the ETAS model from an early aftershock sequence. *Geophysical Research Letters*, **41**(3), 850–857.

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, **31**, 145–155.

Pitkin, J., Manolopoulou, I. and Ross, G. (2018). Bayesian hierarchical modelling of sparse count processes in retail analytics. *arXiv*.

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J. and French, N. P. (2014). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, **13**(1), 67–82.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**(12), 1791–1798.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, **32**(6), 859–866.

Rambaldi, M., Bacry, E. and Lillo, F. (2017). The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, **17**(7), 999–1020.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, **15**(3), 623–642.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G. and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**(4), 366–379.

Ringdal, F. (1975). On the estimation of seismic detection thresholds. *Bulletin of the Seismological Society of America*, **65**(6), 1631–1642.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B*, **39**(2), 172–192.

Rizoiu, M.-A., Lee, Y., Mishra, S. and Xie, L. (2017). A tutorial on Hawkes processes for events in social media. *arXiv*.

Robert, C. P. (2016). The Metropolis-Hastings algorithm. *arXiv*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer.

Ross, G. J. (2021). Bayesian estimation of the ETAS model for earthquake occurrences. *Bulletin of the Seismological Society of America*, **111**(3), 1473–1480.

Ross, G. J. and Kolev, A. A. (2022). Semiparametric Bayesian forecasting of spatiotemporal earthquake occurrences. *The Annals of Applied Statistics*, **16**(4), 2083 – 2100.

Ross, S. M. (1995). *Stochastic Processes*. John Wiley & Sons.

Ruiz, F. J. R., Athey, S. and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, **14**(1), 1 – 27.

Schoenberg, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, **98**(464), 789–795.

Serafini, F., Lindgren, F. and Naylor, M. (2023). Approximation of Bayesian Hawkes process models with inlabru. *Environmetrics*, **34**(5), e2798.

Serre, D. (2002). *Matrices: Theory and Applications*. Springer Science & Business Media, 2nd edition.

Shelton, C., Qin, Z. and Shetty, C. (2018). Hawkes process inference with missing data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press.

Shirota, S. and Gelfand, A. E. (2017). Approximate Bayesian computation and model assessment for repulsive spatial point processes. *Journal of Computational and Graphical Statistics*, **26**(3), 646–657.

Shlomovich, L., Cohen, E. A. K., Adams, N. and Patel, L. (2022). Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, **32**(6), 98.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

Sulem, D., Rivoirard, V. and Rousseau, J. (2021). Bayesian estimation of nonlinear Hawkes process. *arXiv*.

Tan, X., Rao, V. and Neville, J. (2018). Nested CRP with Hawkes-Gaussian processes. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1289–1298. PMLR.

Tetereva, A. (2018). Do financial companies communicate to one another in the news? Application of multivariate Hawkes graphs to uncover Granger causality of financial news. *SSRN (Preprint)*.

Trouleau, W., Etesami, J., Grossglauser, M., Kiyavash, N. and Thiran, P. (2019). Learning Hawkes processes under synchronization noise. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6325–6334. PMLR.

Tucker, D. J., Shand, L. and Lewis, J. R. (2019). Handling missing data in self-exciting point process models. *Spatial Statistics*, **29**, 160–176.

Upadhyay, U., De, A. and Gomez Rodriguez, M. (2018). Deep reinforcement learning of marked temporal point processes. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates Inc.

US Census Bureau (2011). *Section 22. Wholesale and Retail Trade*. Statistical Abstract of the United States. Government Printing Office.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, **103**(482), 614–624.

Vihola, M. and Franks, J. (2020). On the use of approximate Bayesian computation Markov chain Monte Carlo with inflated tolerance and post-correction. *Biometrika*, **107**(2), 381–395.

Williams, P. J., Hooten, M. B., Womble, J. N. and Bower, M. R. (2017). Estimating occupancy and abundance using aerial images with imperfect detection. *Methods in Ecology and Evolution*, **8**(12), 1679–1689.

Xu, H., Luo, D. and Zha, H. (2017). Learning Hawkes processes from short doubly-censored event sequences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3831–3840. PMLR.

Zuo, S., Jiang, H., Li, Z., Zhao, T. and Zha, H. (2020). Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11692–11702. PMLR.

Appendix A

Additional Examples and Material for ABC-Hawkes

This appendix gives additional examples for the cases examined in Chapter 3. In addition, it also gives an introduction to semiautomatic ABC and how it could be used for Hawkes processes when data is missing.

A.1 Additional Examples

This section provides additional examples on other data sets for all simulation studies undertaken in Chapter 3.

A.1.1 No Distortion: Recover Posterior

In this section, we provide additional examples for Section 3.4.1, where there is no data distortion. This shows that ABC-Hawkes can recover the posterior distributions well when the data is complete.

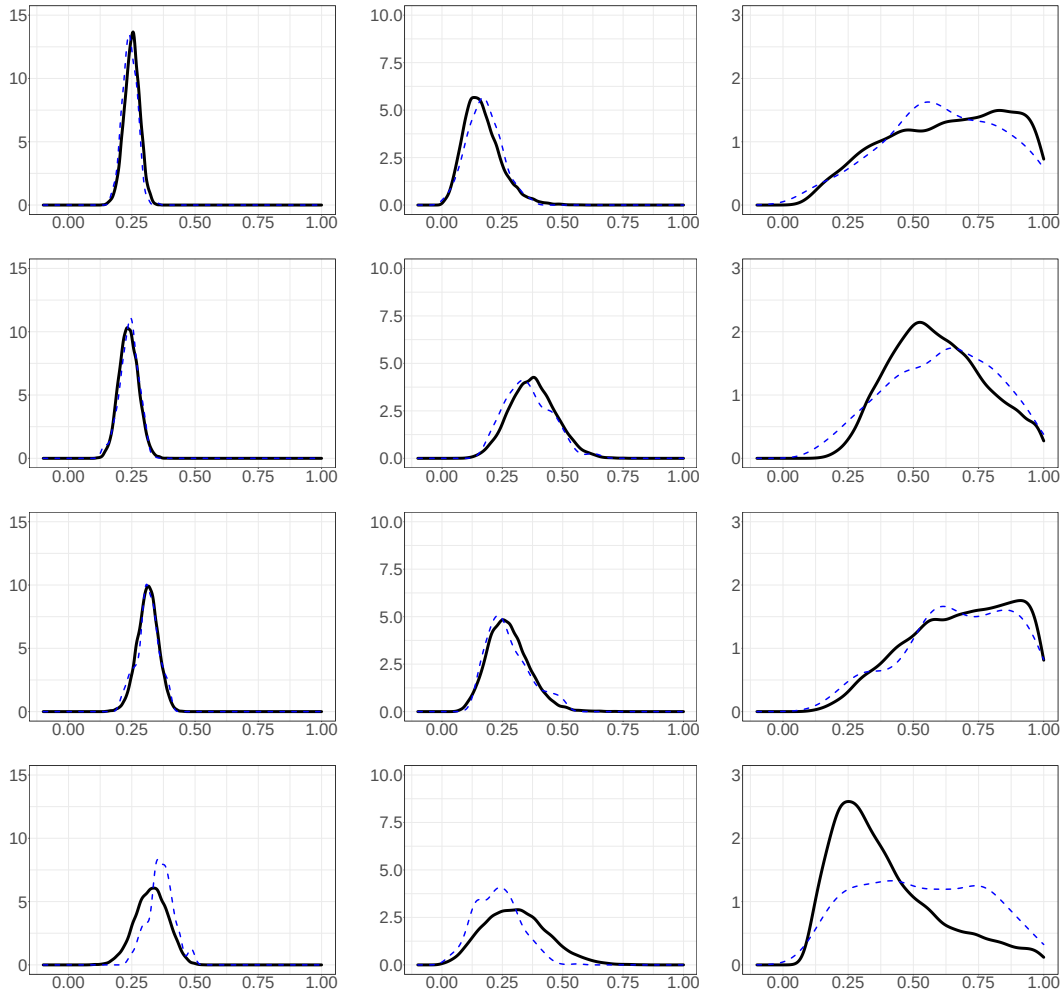


Figure A.1: Undisturbed data posterior distributions for μ (**left**), K (**middle**), and β (**right**). The solid curve shows the true posterior estimated by samples generated from Stan, dashed represents ABC-Hawkes. All estimates are based on the complete, undistorted data sets.

A.1.2 No Distortion: Comparison Ertekin

In this section, we compare ABC-Hawkes to the proposed method by Ertekin *et al.* (2015), as shown in Section 3.4.1. As discussed there, we assume μ to be known.

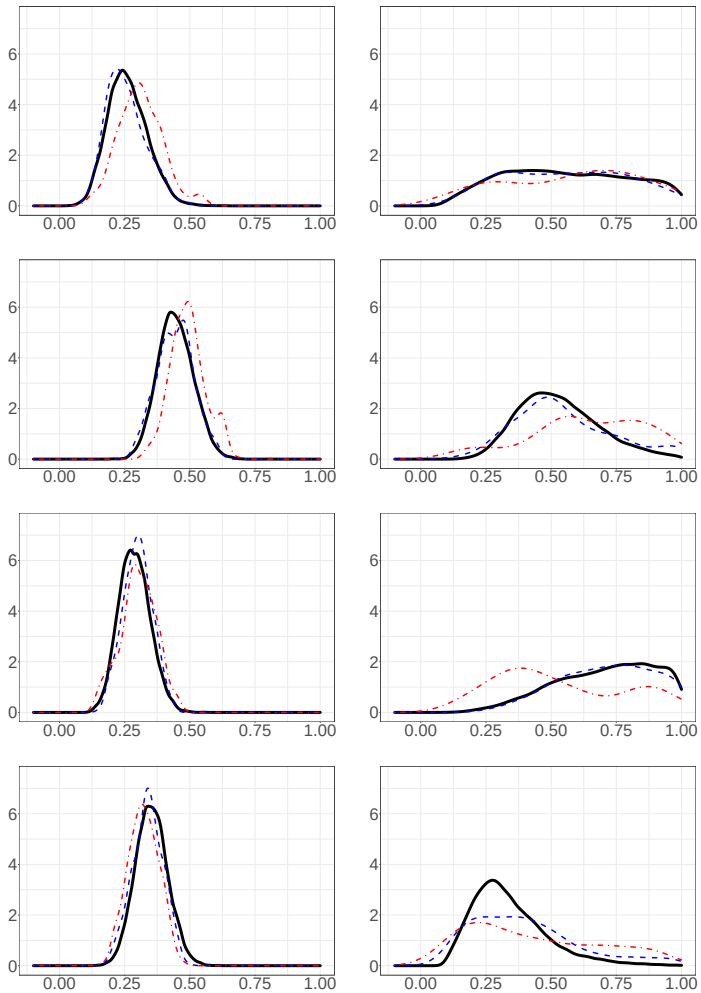


Figure A.2: Undisturbed data posterior distributions for μ (**left**), K (**middle**), and β (**right**). The solid curve shows the true posterior estimated by samples generated from Stan, dashed represents ABC-Hawkes, dot-dashed uses the summary statistics from Ertekin *et al.* (2015). All estimates are based on the complete, undistorted data sets.

A.1.3 Constant Deletion, ξ Known

This section gives additional simulations for Section 3.4.3, where events are observed with a known, constant probability a.

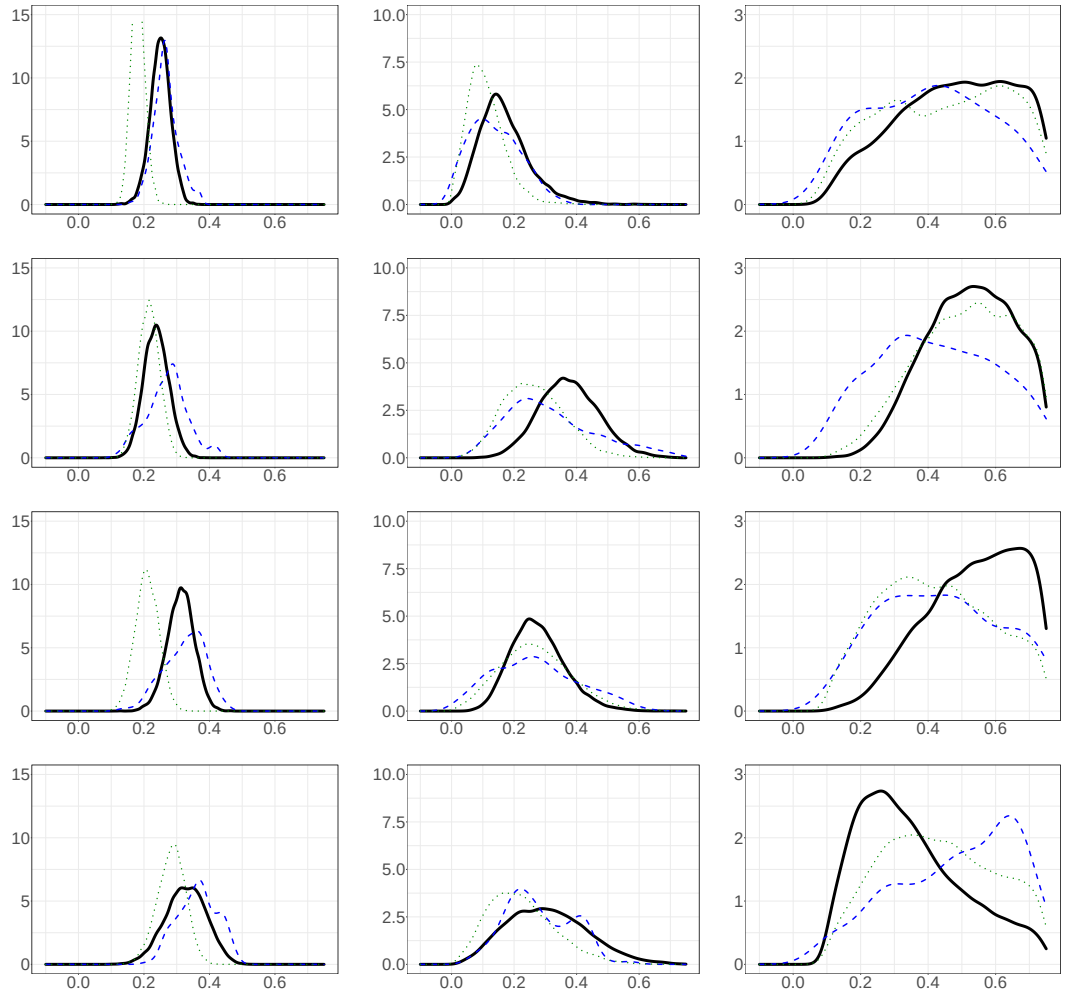


Figure A.3: Data with known constant deletion posterior distributions for μ (**left**), K (**middle**), and β (**right**). The solid curve shows the true posterior estimated by samples generated from Stan, dashed represents ABC-Hawkes, the naive approach using only the incomplete data is dotted.

A.1.4 Constant Deletion, ξ Unknown

Here we give additional examples for a constant deletion when the probability of observing an event is unknown, as shown in Section 3.4.3. As described, the success of the posterior recovery depends on the parameters due to identifiability issues. We provide both examples here.

More Successful Posterior Recovery

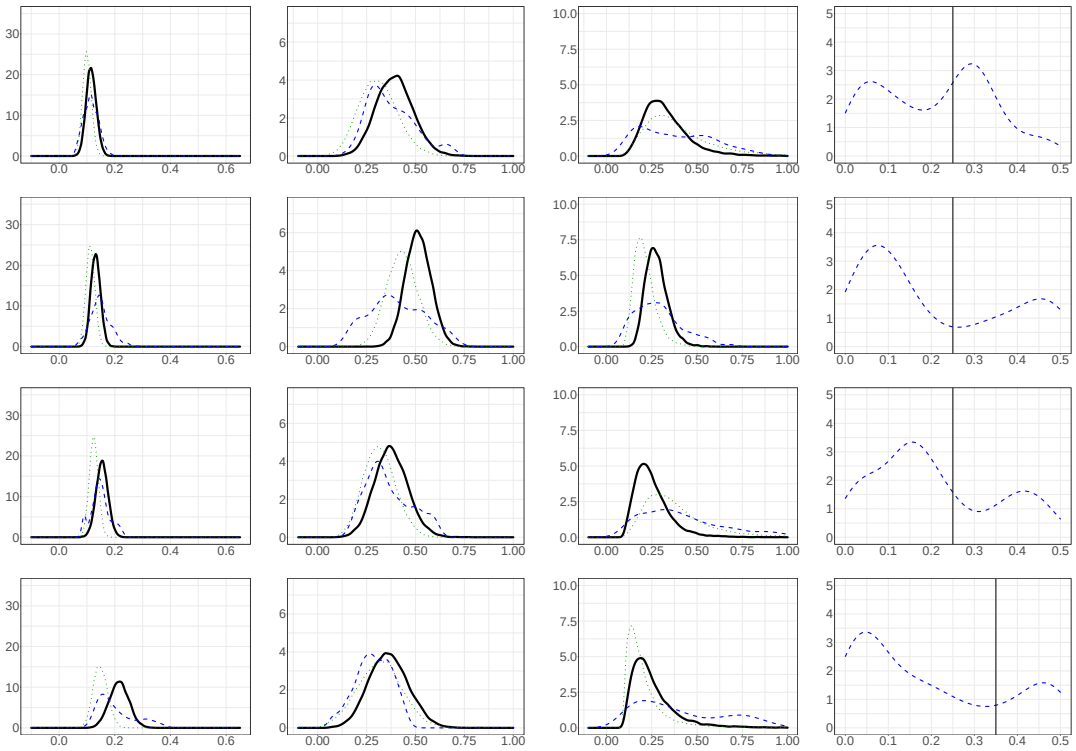


Figure A.4: Data with unknown constant deletion posterior distributions for μ (**first**), K (**second**), β (**third**), and $1 - a$ (**fourth**). The solid curve shows the true posterior estimated by samples generated from Stan, dashed represents ABC-Hawkes, the naive approach using only the incomplete data is dotted. The vertical line in the fourth plot represents the true value.

Less Successful Posterior Recovery

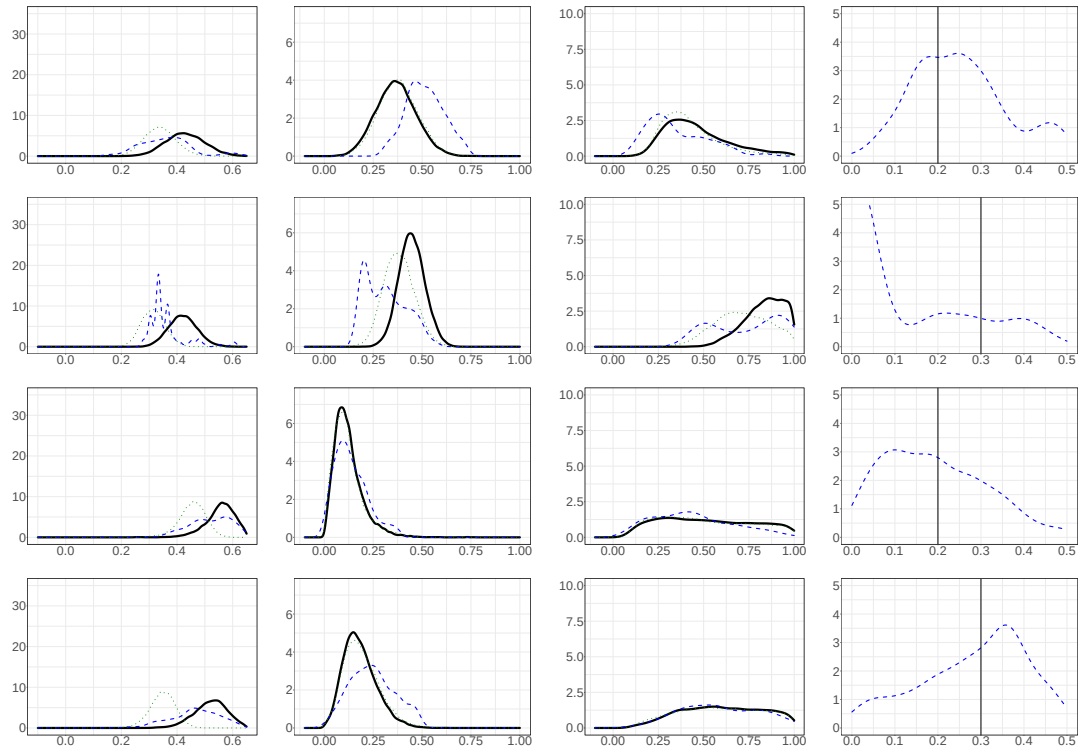


Figure A.5: Data with unknown constant deletion posterior distributions for μ (**first**), K (**second**), β (**third**), and $1 - a$ (**fourth**). The solid curve shows the true posterior estimated by samples generated from Stan, dashed represents ABC-Hawkes, the naive approach using only the incomplete data is dotted. The vertical line in the fourth plot represents the true value.

A.2 Semiautomatic ABC

It has been suggested to us that semiautomatic ABC might provide a good performance for our problem setting in Chapter 3. In this appendix, we discuss a semiautomatic setup for Hawkes processes under data distortion. Despite our best efforts, the semiautomatic approach does not improve on the ABC-Hawkes results, as shown below.

Fearnhead and Prangle (2012) introduce a semiautomatic approach to selecting summary statistics. In essence, they use weighted auxiliary summary statistics as estimates of the posterior mean. Semiautomatic ABC nevertheless requires users to specify auxiliary summary statistics, but the method is described as more robust as more or less relevant statistics are weighted appropriately. It relies on this large number of M auxiliary summary statistics to together contain enough information to estimate the parameter. For example, when covariates available Fearnhead and Prangle (2012) recommend to use different powers of the covariates as auxiliary statistics. This approach has been successfully used for a number of applications, such as bacteria pathogens (Prangle *et al.*, 2014) or component reliability (McCullough and Ebrahimi, 2018).

The semiautomatic workflow for a one-dimensional parameter θ is as follows, for a multivariate parameter vector θ the workflow is followed for each entry of the vector separately.

1. **Pilot run:** A pilot run determines regions of the posterior distribution for θ with non-negligible mass.

2. **Weight Sampling:** Parameters $\phi = (\phi_1 \dots \phi_N)$ are drawn from this areas and subsequently used to simulate data sets $(X_1 \dots X_N)$.

3. **Weight Calculation:** For each $i = 1 \dots N$ a large number of auxiliary summary statistics $\overline{S_i} = \left(\overline{S_{1i}}(X_i) \dots \overline{S_{Mi}}(X_i)\right)$ is calculated and $N \times M$ matrix $\overline{S}$ is constructed such that $\overline{S_i}$ is in row i. Then, the standardised linear regression parameters b in

$$\phi = b_0 + \overline{S}\, b + \varepsilon \tag{A.1}$$

   are estimated where ε is random error.

4. **ABC:** Run ABC with the single summary statistic

$$S(Y^{(j)}) = \left(\overline{S_{(j)1}}(Y^{(j)}) \dots \overline{S_{(j)M}}(Y^{(j)})\right)\, b, \tag{A.2}$$

   where $S(Y^{(j)})$ are linear combination weighted by b of all M auxiliary summary statistics.

We have implemented the semiautomatic according to Algorithm 3, which serves as a comparison to Section A.2.2. Details on the chosen auxiliary statistics can be found in Section A.2.1.

Algorithm 3 Semiautomatic ABC (for Hawkes with known ξ)

Pilot run: determine area of non-negligible posterior mass $\tilde{\pi}(\theta)$
Weight Sampling:
for $i = 1$ **to** N **do**
  $\phi^{(i)} \sim \tilde{\pi}(\cdot)$
  $W^{(i)} \sim p(\cdot \mid \phi^{(i)})$, where $p(\cdot)$ simulates from a Hawkes process
  distort data $W^{(i)}$ according to $h(W^{(i)} \mid \xi)$, the resulting data set is $X^{(i)}$
end for
Weight Calculation:
for $i = 1$ **to** N **do**
  calculate $\overline{S_i} = \left(\overline{S_{1i}}(X_i) \dots \overline{S_{Mi}}(X_i)\right)$
end for
for $p = 1$ **to** P **do**
  estimate the regression parameters vector b_p from $\phi_{ip} = b_0 + \overline{S_i}\, b_p + \varepsilon$
  calculate $sd_p = sd(\overline{S_i}\, b_p)$
end for
ABC:
Initialise $\theta^{(0)}$
for $j = 1$ **to** J **do**
  $\theta^* \sim q(\cdot \mid \theta^{(j-1)})$
  $Z^* \sim p(\cdot \mid \theta^*)$, where $p(\cdot)$ simulates from a Hawkes process
  distort data Z^* according to $h(Z^* \mid \xi)$, the resulting data set is Y^*
  calculate $\overline{S}(Y^*) = \left(\overline{S_1}(Y^*) \dots \overline{S_M}(Y^*)\right)$
  if $\left(\sum_{p=1}^{P} \left[\left(\overline{S}(Y^*) - \overline{S}(Y)\right) b_p / sd_p\right]^2\right)^{-2} < \epsilon$ **then**
    With probability $\min\left\{1, \frac{q(\theta^{(j-1)} \mid \theta^*)\pi(\theta^*)}{q(\theta^* \mid \theta^{(j-1)})\pi(\theta^{(j-1)})}\right\}$ set $\theta^{(j)} = \theta^*$
  else
    Set $\theta^{(j)} = \theta^{(j-1)}$
  end if
end for
Output: $(\theta^{(1)}, \dots, \theta^{(J)})$

A.2.1 Auxiliary Summary Statistics

Despite its name semiautomatic ABC not all aspects are autonomous and the method still requires the definition of a large number of auxiliary summary statistics. The particular data structure, as well as the lack of covariates, required us to design a new set of auxiliary summary statistics. For semiautomatic ABC, the following $M = 92$ are used, where the event time differences are defined as $\Delta_i = t_i - t_{i-1}$ for $i = 1 \dots N - 1$.

- Number of events
- Log(number of events)
- Ripley's K (Ripley, 1977) of window size $(1, 2 \dots 9, 10)$
- The quantiles $(0.05, 0.1 \dots 0.9, 0.95)$ of the interevent times Δ
- The quantiles $(0.05, 0.1 \dots 0.9, 0.95)$ of the interevent times Δ divided by the average of Δ
- Median of Δ divided by the quantiles $(0.1, 0.25, 0.75, 0.9)$ of Δ
- Average of Δ that lie above the $(0.05, 0.1 \dots 0.9, 0.95)$ quantile of Δ
- Average of Δ that lie below the $(0.05, 0.1 \dots 0.9, 0.95)$ quantile of Δ

A.2.2 Examples

We now examine how the semiautomatic approach can be employed to estimate Hawkes processes and compare it to ABC-Hawkes with seven summary statistics in a variety of scenarios. First, we give an example where the semiautomatic ABC can successfully recover the true posterior on complete data. Second, we show and example where semiautomatic ABC is not able to recover the posterior distribution to a satisfactory degree. We finish with a discussion regarding the employability of the semiautomatic approach.

No distortion

We implemented the semiautomatic ABC along the lines of Algorithm 3 for complete data. Figure A.6 gives the posterior distributions, which show a successful recovery of the correct posterior distribution, similar to ABC-Hawkes with seven summary statistics.

Twitter

As a second example, we examine the Twitter example from Section 3.4.2, where the gap is known. First, we use semiautomatic ABC as above with a simple pilot run to determine $\tilde{\pi}(\cdot)$, the non-negligible areas of posterior mass. Note that these areas identified in the pilot run are only used in the subsequent weight sampling step (as outlined in Algorithm 3) and do not contribute otherwise. As

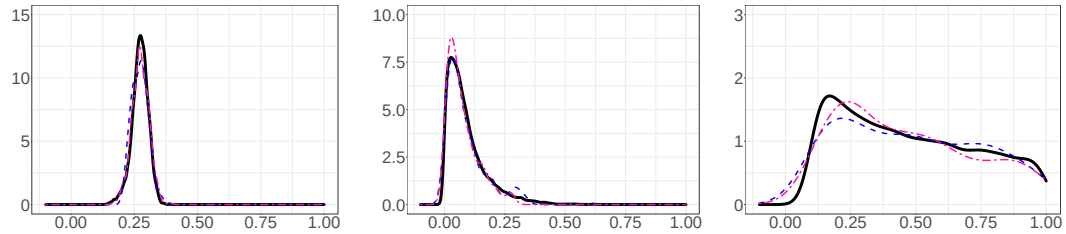


Figure A.6: Undisturbed data posterior distributions for μ (**left**), K (**middle**), and β (**right**). The black/solid curve shows the true posterior estimated by samples generated from Stan, blue/dashed represents ABC-Hawkes with seven summary statistics, pink/two-dashed is semiautomatic. All estimates are based on the complete, undistorted data sets.

suggested by Fearnhead and Prangle (2012), we choose $\tilde{\pi}(\cdot)$ to be a hypercube in the parameter space.

Figure A.7 provides the posterior distributions for this data set. It is evident that semiautomatic cannot recover the mean and standard deviation, let alone the full posterior distributions.

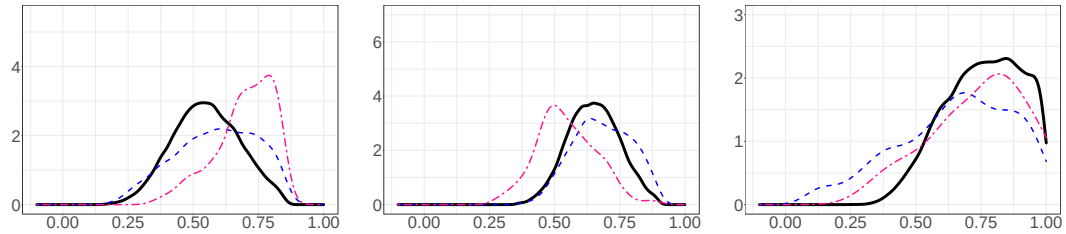


Figure A.7: Twitter example with known gap posterior distributions for μ (**left**), K (**middle**), and β (**right**) with simple $\tilde{\pi}(\cdot)$. The black/solid curve shows the true posterior estimated by samples generated from Stan, blue/dashed represents ABC-Hawkes with seven summary statistics, pink/two-dashed is semiautomatic.

There is the possibility that this is due to the fact that the trial run was not elaborate enough. In general, we found that the pilot run of the semiautomatic algorithm, which is described as "optional" by Fearnhead and Prangle (2012), can improve the results of the algorithm. To enable optimal performance of the semiautomatic ABC, we now use the posterior samples from ABC-Hawkes with seven summary statistics as $\tilde{\pi}(\cdot)$. This is a rather costly and informative trial run, but wanted to give the semiautomatic approach all possible resources to succeed. For reference, Fearnhead and Prangle (2012) suggest that the trial run should take about a quarter of all computational time, whereas it takes up closer to a half in this case. Nevertheless, we deem this a necessary step to use the ABC-Hawkes with seven summary statistics as a pilot run if semiautomatic ABC is to be used at all.

Hence, for the next example, we again use the Twitter data with a known gap, but this time provide the samples from ABC-Hawkes to the semiautomatic approach as a trial run. Despite this considerable "help", the semiautomatic ABC

is not able to recover the posterior distributions, as showcased in Figure A.8. It clearly shows that the semiautomatic approach does not match – and certainly does not exceed – the performance of ABC-Hawkes with seven summary statistics, even with a more optimal pilot run.

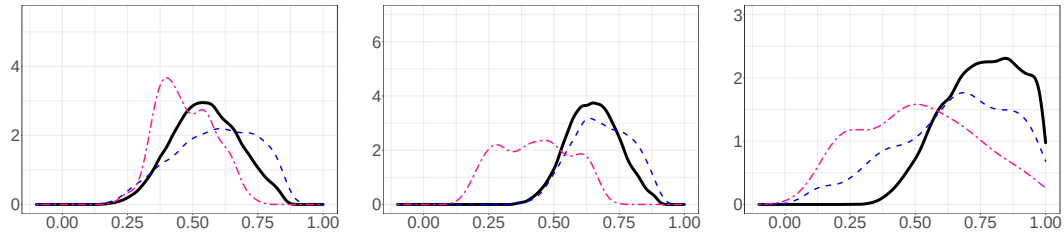


Figure A.8: Twitter example with known gap posterior distributions for μ (**left**), K (**middle**), and β (**right**) with ABC-Hawkes as pilot run. The black/solid curve shows the true posterior estimated by samples generated from Stan, blue/dashed represents ABC-Hawkes, pink/two-dashed is semiautomatic.

A.2.3 Discussion

It has been a persistent issue in our experiments that the semiautomatic approach produced less satisfactory results compared to the ABC-Hawkes with seven summary statistics, even when samples from ABC-Hawkes are used as a pilot run. As a minor point, the semiautomatic approach also has a noticeably longer run time, as is requires the ABC-Hawkes as a trial run for better performance. It also takes longer on each iteration of the MCMC as substantially more summary statistics need to be computed (92 as opposed to 7 for ABC-Hawkes). Despite our best efforts to improve the estimation procedure of Hawkes processes with missing data through semiautomatic ABC, the semiautomatic results were continuously less satisfactory and hence we stick to our originally proposed approach of ABC-Hawkes with seven summary statistics.

Appendix B

Extended Model Summaries for Group Chat Application

This appendix contains additional material for Chapter 6. Due to the large number of parameters only a subset of plots was shown in the main text. All posterior distributions, as well as more detailed summary statistics, are displayed here.

B.1 Posterior Distributions

This section contains the posterior distributions for all parameters of both the Ancestor Hawkes and the classic Hawkes for the group chat data described in Section 6.5.

B.1.1 Background Scalar

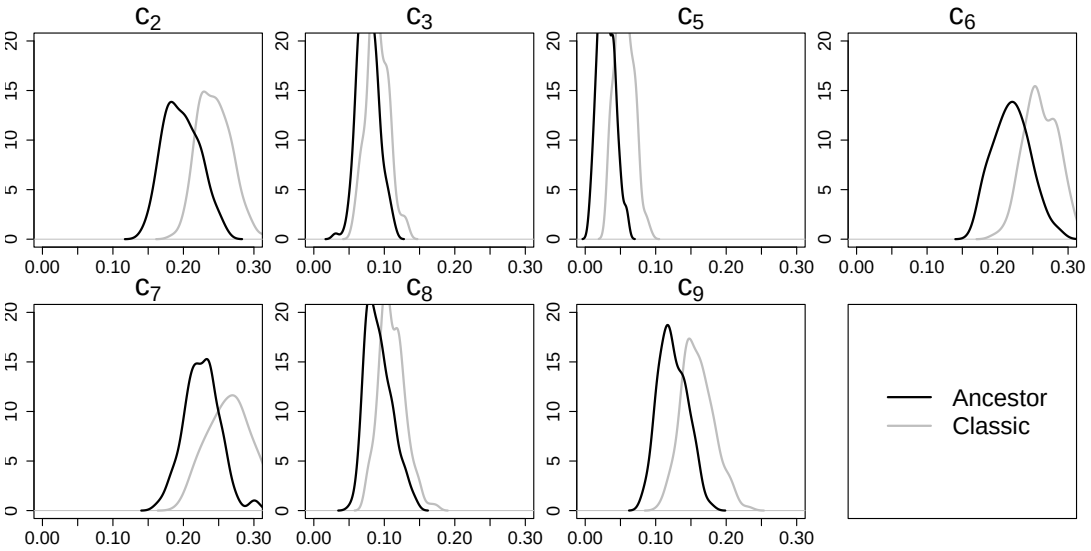
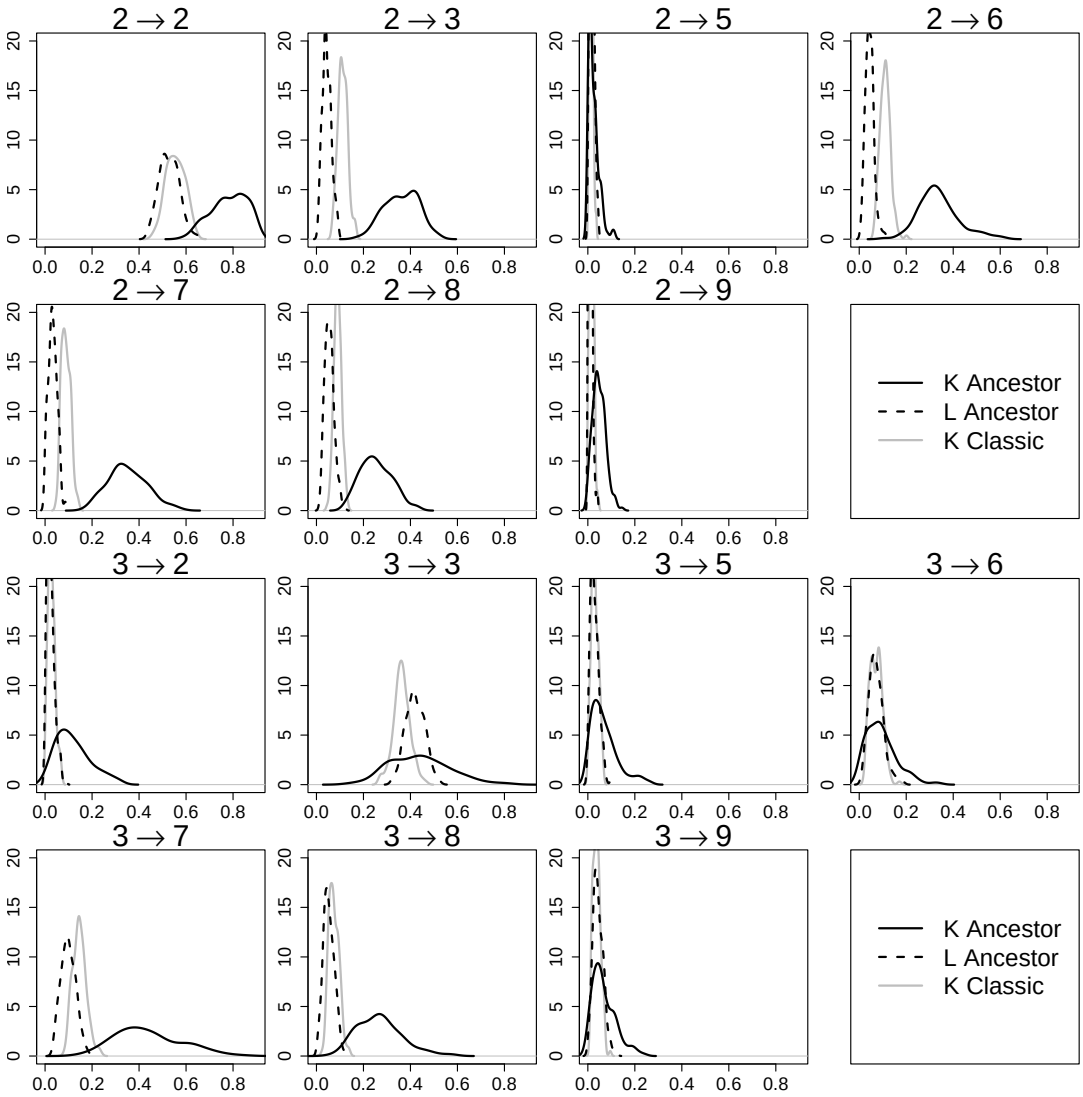
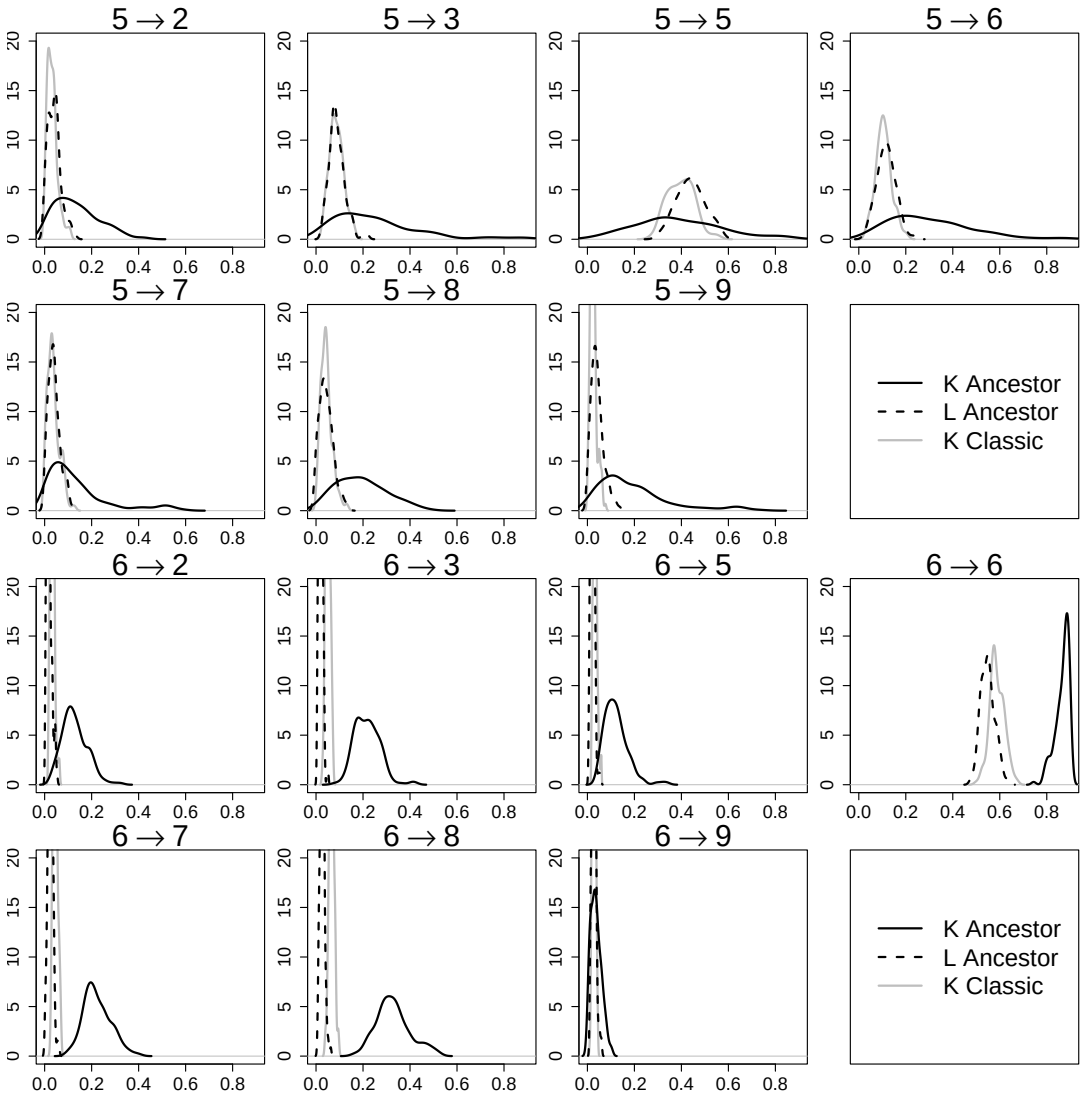
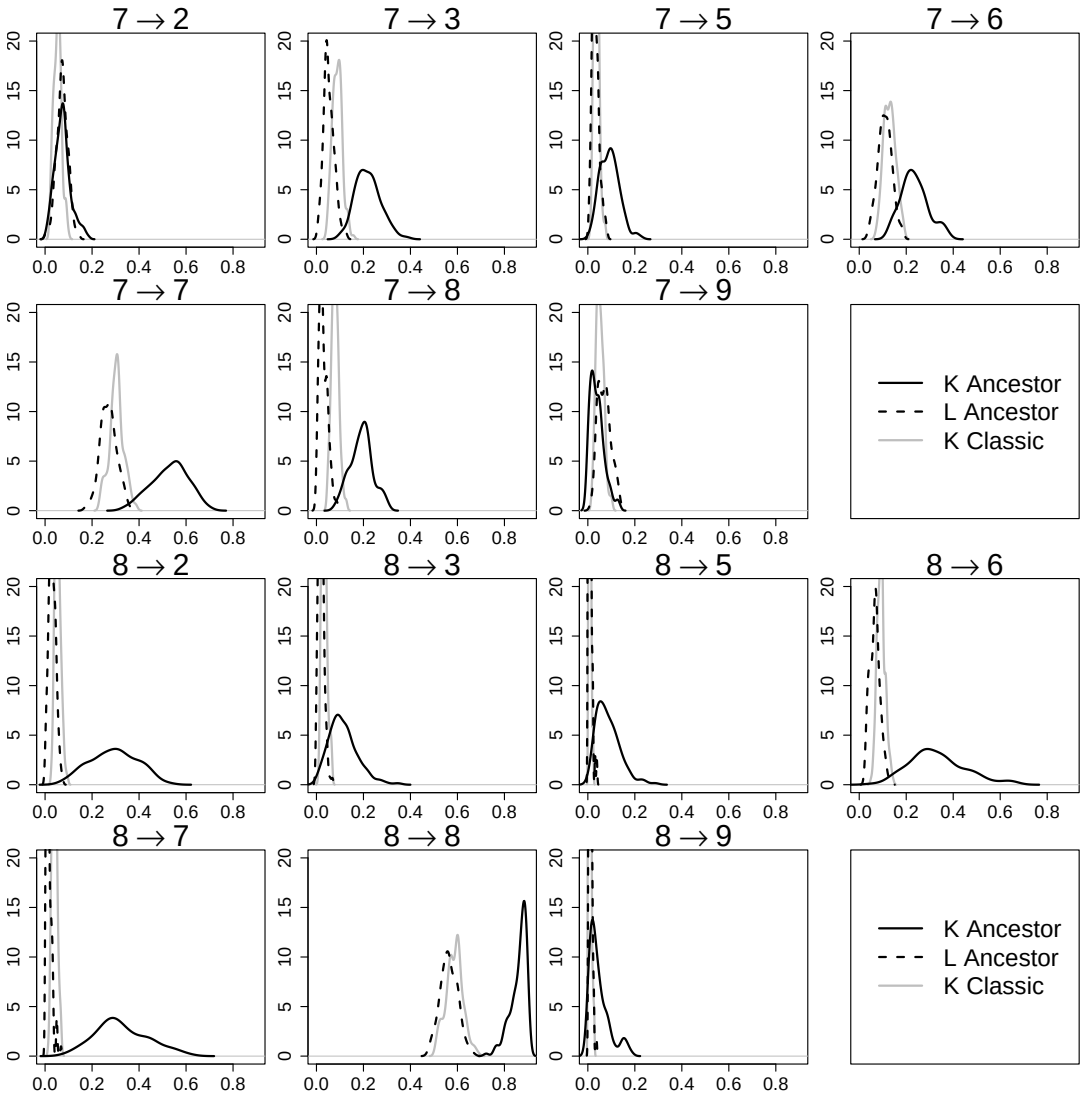


Figure B.1: Posterior distribution estimates for the scalar of the background rate for each participant. Participant 1 and 4 were excluded due to their low engagement. Black line gives posterior distribution in Ancestor Hawkes model, grey line shows posterior distribution for the classic Hawkes.

B.1.2 Influence Magnitude




5 → 2
5 → 3
5 → 5
5 → 6
5 → 7
5 → 8
5 → 9
6 → 2
6 → 3
6 → 5
6 → 6
6 → 7
6 → 8
6 → 9
K Ancestor
L Ancestor
K Classic
0 5 10 15 20
0.0 0.2 0.4 0.6 0.8


7 → 2
7 → 3
7 → 5
7 → 6
7 → 7
7 → 8
7 → 9
8 → 2
8 → 3
8 → 5
8 → 6
8 → 7
8 → 8
8 → 9
K Ancestor
L Ancestor
K Classic
K Ancestor
L Ancestor
K Classic

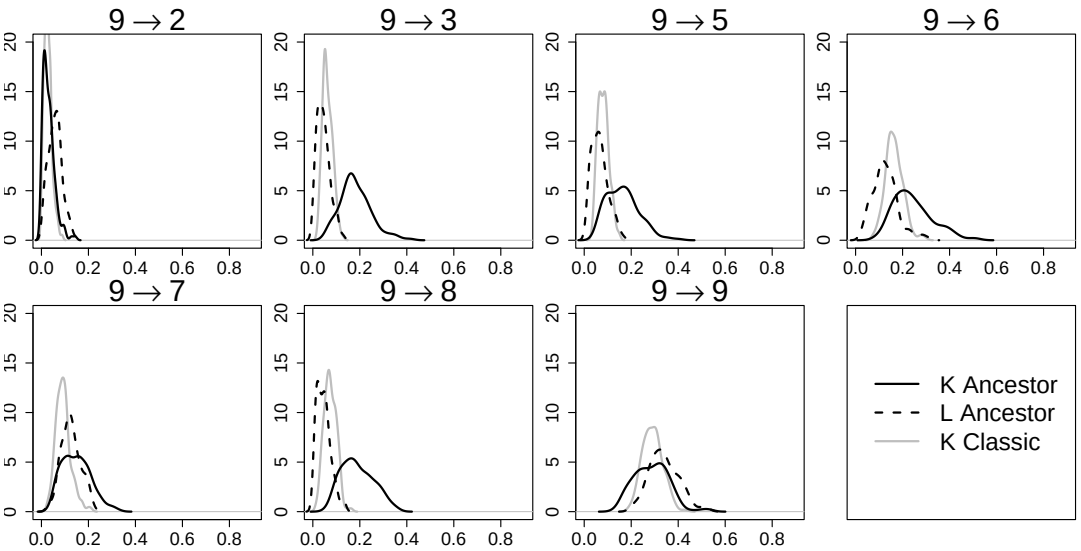


Figure B.2: Posterior distribution estimates for the influence magnitude for all participants. Participant 1 and 4 were excluded due to their low engagement. Black solid line gives posterior distribution for **K** in Ancestor Hawkes model and black dashed line represents posterior estimates for **L** in Ancestor Hawkes, grey line shows posterior distribution of **K** in the classic Hawkes.

B.1.3 Influence Kernel

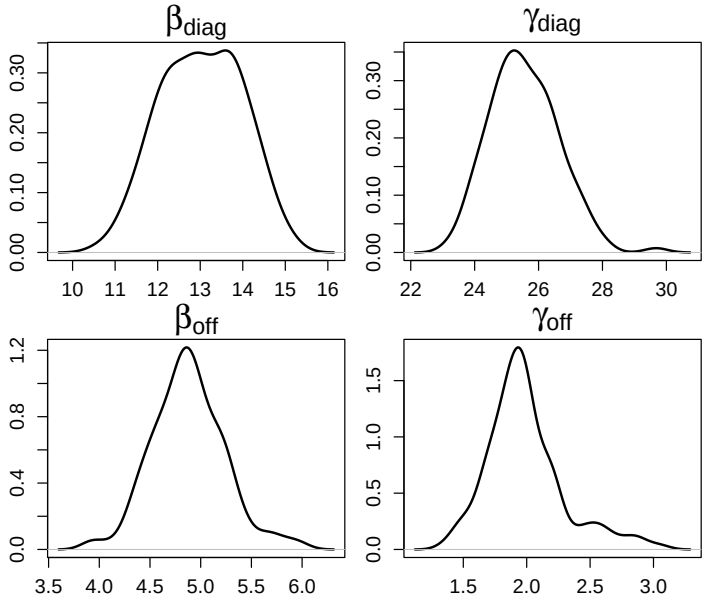


Figure B.3: Posterior Distributions for the influence kernel parameters in the Ancestor Hawkes model. Note the different axes on each sub-plot.

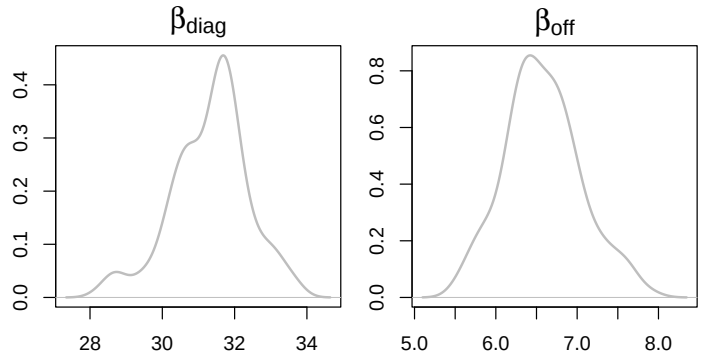


Figure B.4: Posterior Distributions for the influence kernel parameters in the classic Hawkes model. Note the different axes on each sub-plot.

B.2 Summary Statistics

This section gives additional plots for summary statistics as discussed in Section 6.5.4. Each summary statistic is provided across the whole data set, as well as for each dimension separately.

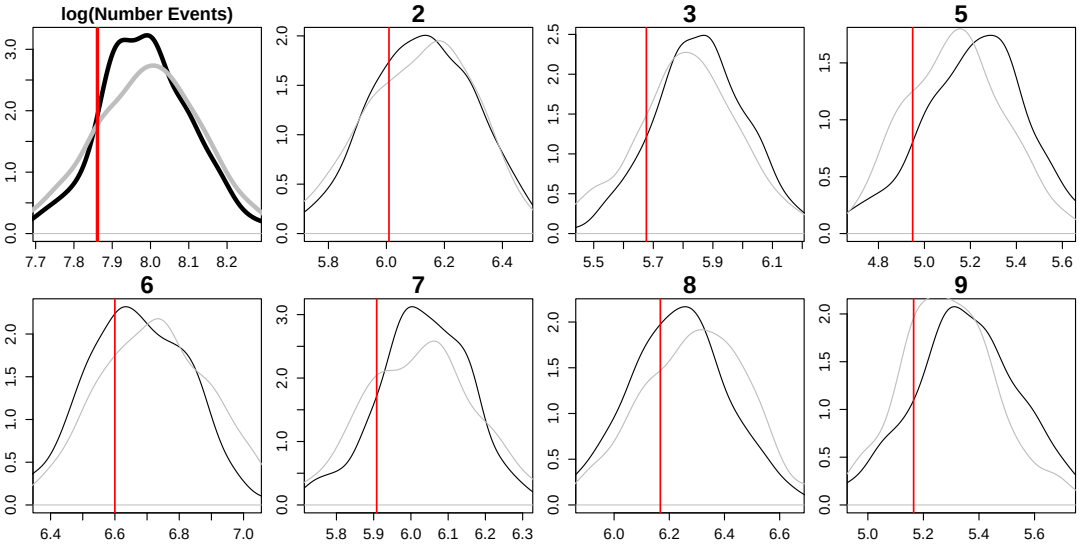


Figure B.5: Comparison of distribution of summary statistic which gives the logarithm of the number of events. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

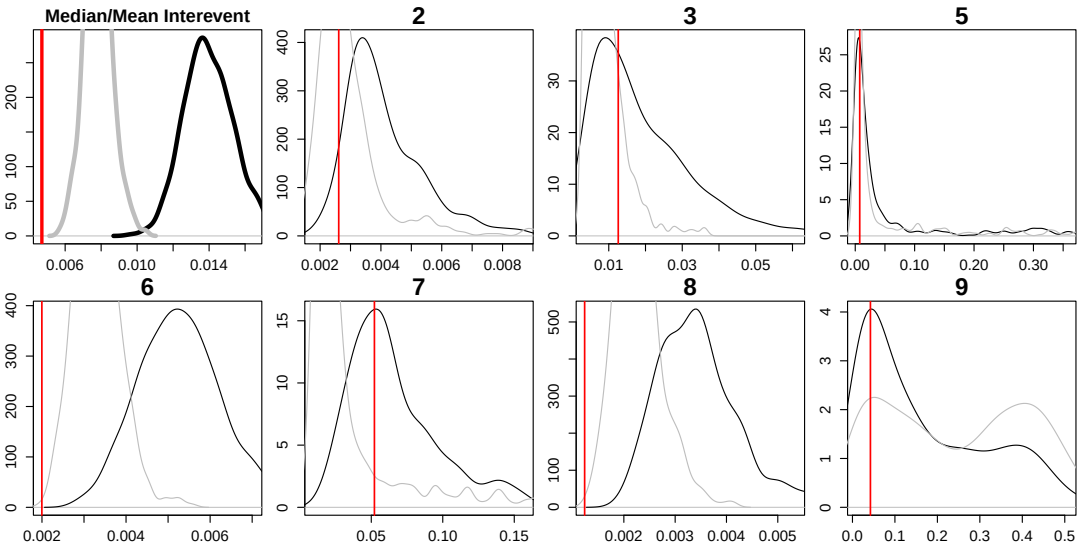


Figure B.6: Comparison of distribution of summary statistic which caluculates the median of interevent times over the mean of interevent times. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

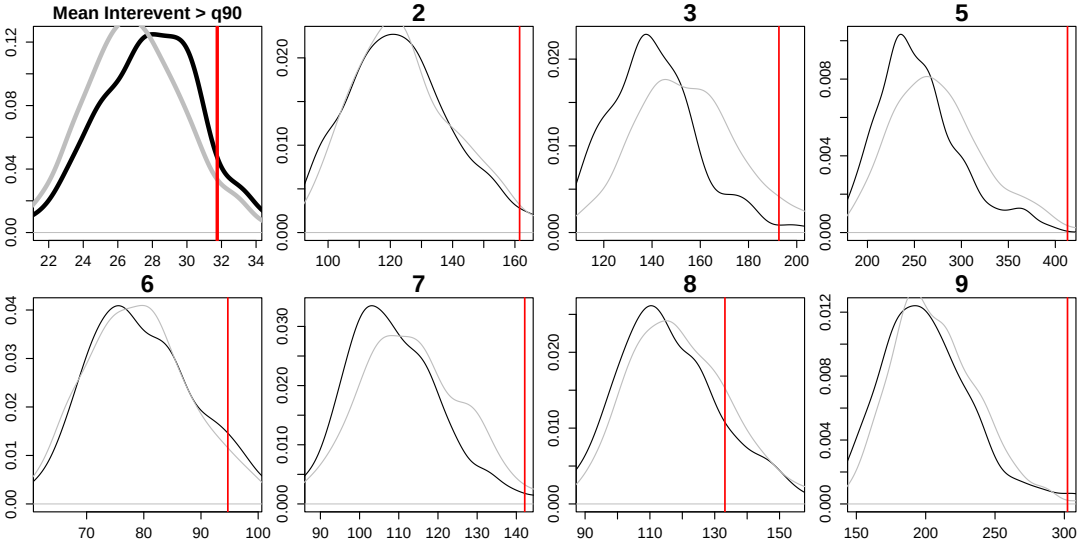


Figure B.7: Comparison of distribution of summary statistic which gives the mean interevent times for those that lie above the 90% quantile. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

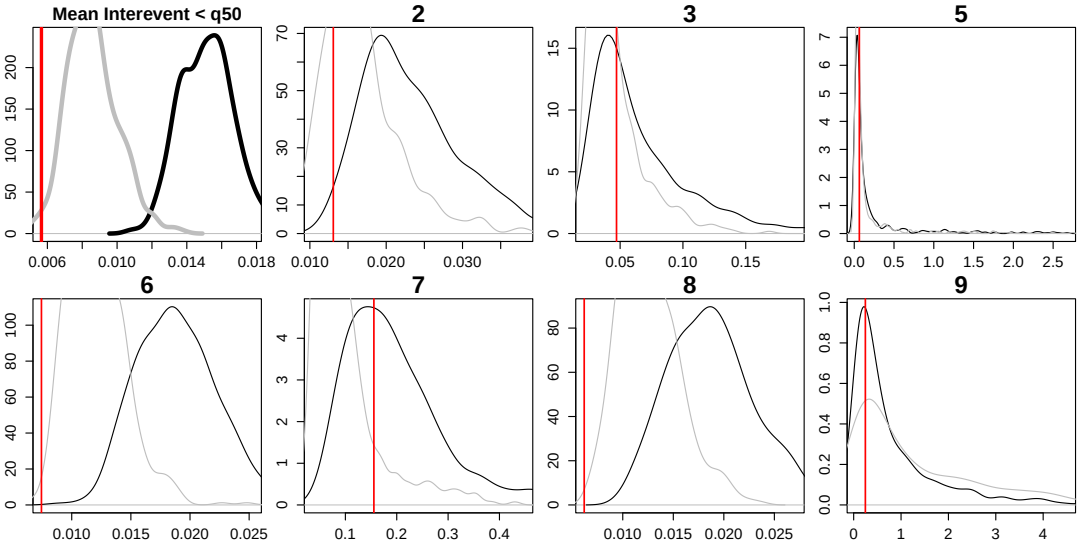


Figure B.8: Comparison of distribution of summary statistic which gives the mean interevent times for those that lie below their median. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

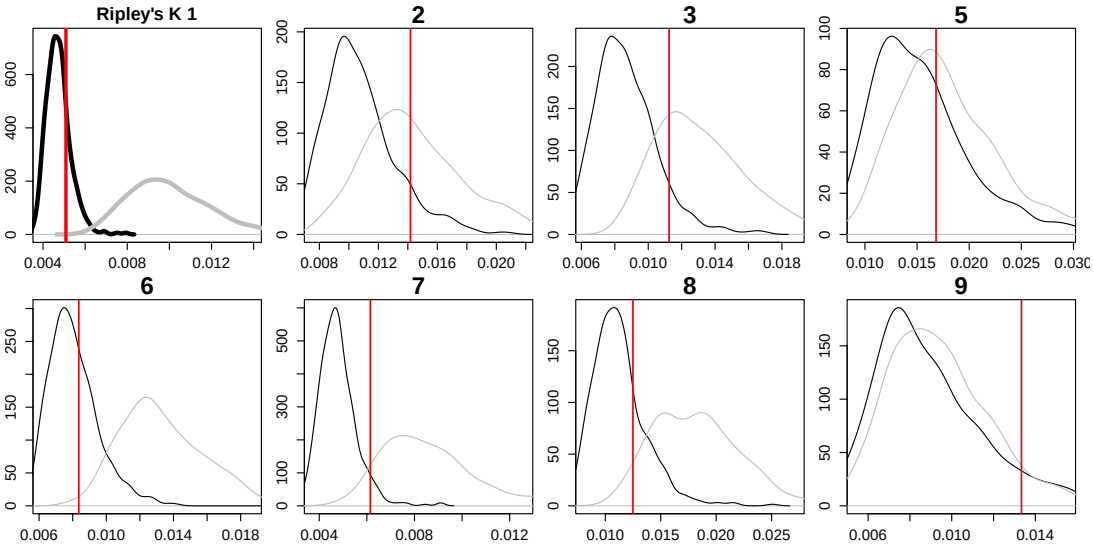


Figure B.9: Comparison of distribution of summary statistic Ripley's K with a window size of 1. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

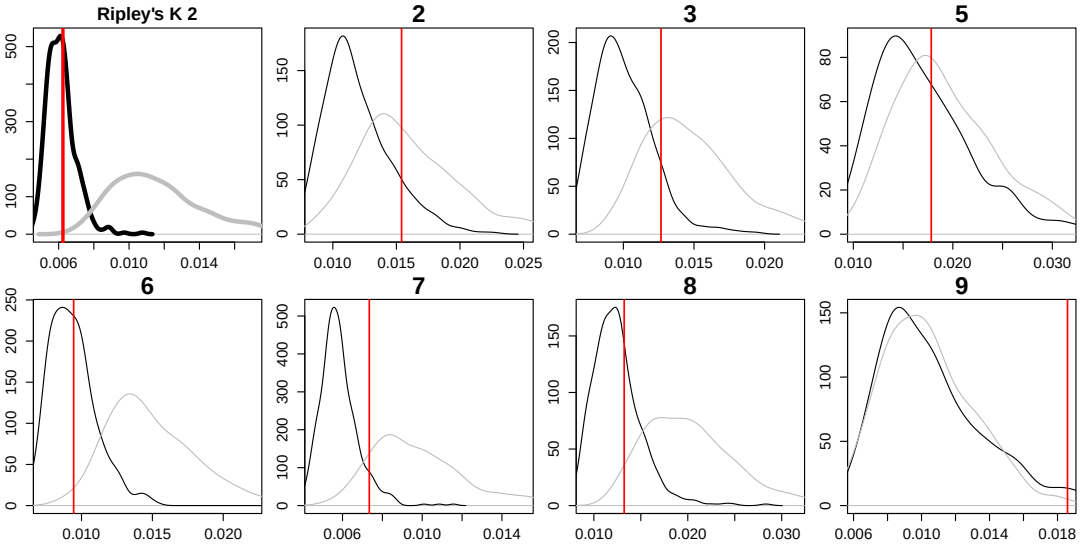


Figure B.10: Comparison of distribution of summary statistic Ripley's K with a window size of 2. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

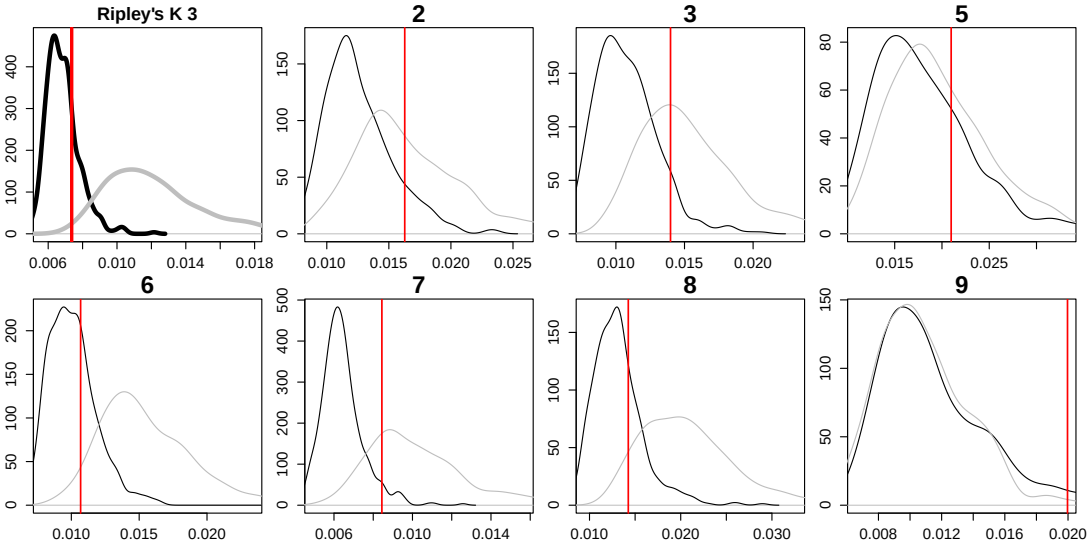


Figure B.11: Comparison of distribution of summary statistic Ripley's K with a window size of 3. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

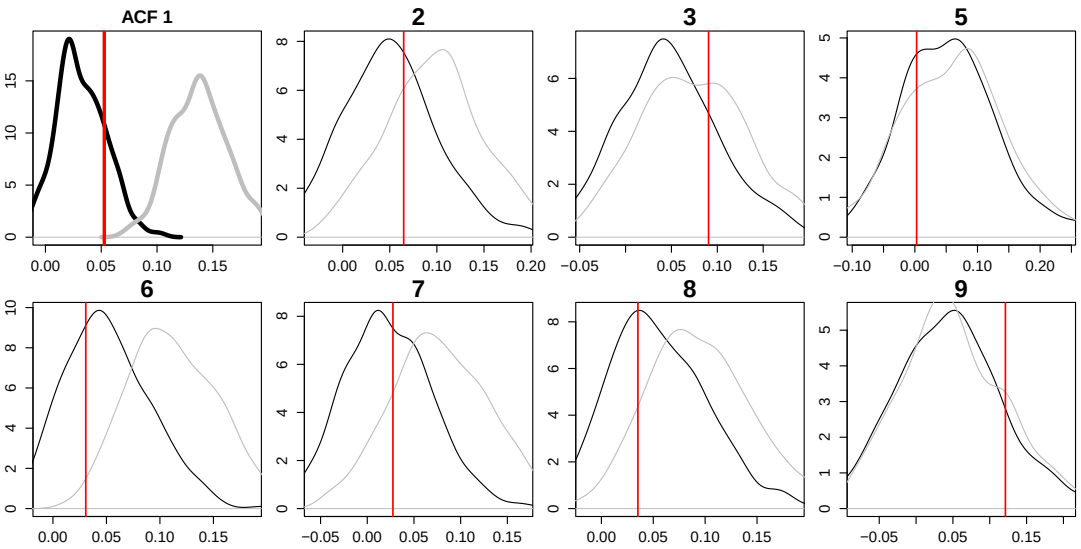


Figure B.12: Comparison of distribution of the autocorrelation of interevent times with lag 1. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

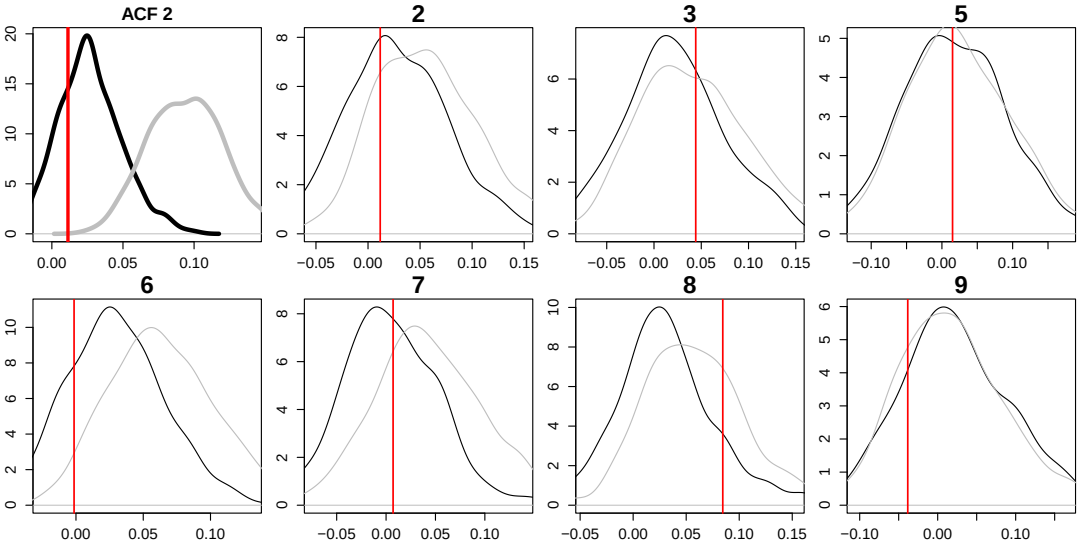


Figure B.13: Comparison of distribution of the autocorrelation of interevent times with lag 2. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.

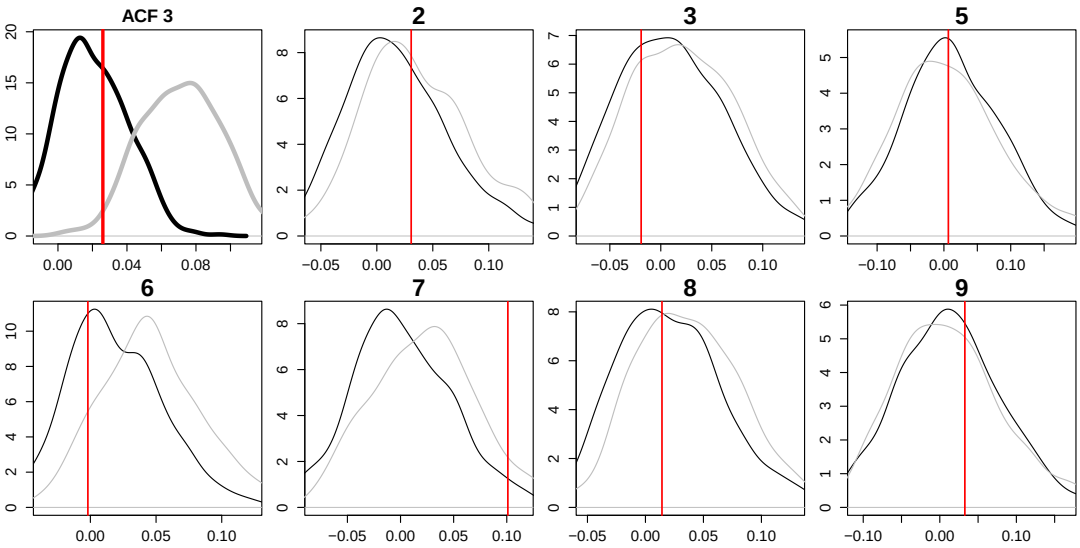


Figure B.14: Comparison of distribution of the autocorrelation of interevent times with lag 3. Black line shows distribution when data is simulated from the Ancestor Hawkes model. Grey line gives the distribution for the classic Hawkes. Red vertical bar indicates the true value from the observed group chat data. **Top left** plot gives the statistic calculated across the whole data set, the other plots are for each participant separately.